Innovative methodology for Bayesian hierarchical modelling, with applications in biology and toxicology

Witold Wiecek

September 2019

Acknowledgements

Research presented in this thesis happened simultaneously with my work at Analytica Laser/Certara. Completing this work would not be possible without the support and contributions of my colleagues. In particular, I thank Billy Amzal for his trust, advice, and his ambition in supporting innovative Bayesian research. Without the flexibility that I was allowed, I would not even consider working on a thesis. Thanks also go to my colleagues Helene Karcher for always helping me find time to do research and Nadia Quignot for patiently teaching me what little I know about toxicology.

Over the last three years, collaborating with my co-supervisor Dr Frédéric Bois has been both my great pleasure and an inspiration to improve my skills. Frédéric, thank you for your energy and for finding time to work together, often at very odd hours and in-between other projects. (And sorry for all the bugs.) I also want to thank Dr Ghislaine Gayraud at UTC Compiegne for her great statistical expertise that made the Bayesian networks article possible.

I am indebted to Prof. Marc Aerts, my Hasselt university co-supervisor, for giving me the opportunity to bring my research to Hasselt and the gift of his time and advice. I couldn't have wished for a more welcoming and less stressful atmosphere to bring this work to completion and I can only hope my association with the research community at UHasselt will continue past this thesis.

I want to thank other researchers who collaborated with me. In particular, Dr Jean-Lou Dorne at the European Food Safety Authority, for working very hard with me so that we can publish the research on chemical risk assessment. Although Dr Jose Abrahantes was not involved in the work included here, it was his advice that helped me lock down the contents of this thesis and bring it to completion.

Most importantly, I thank my loved ones and friends in London and Kraków for cheering me on, especially through many weekends and evenings. Working on this thesis was definitely a case of bringing the work home with me and I don't think it would be possible to complete this without the fantastic support network that I had around me. Commiserations to those of you, who asked what my research was about and then had to sit through the whole explanation.

List of publications

Wiecek, Witold, Frederic Bois, and Ghislaine Gayraud. "Structure Learning of Bayesian Networks Involving Cyclic Structures", 2019 (in submission).

Quignot, Nadia, Witold Wiecek, Billy Amzal, and Jean-Lou Dorne. "The Yin–Yang of CYP3A4: A Bayesian Meta-Analysis to Quantify Inhibition and Induction of CYP3A4 Metabolism in Humans and Refine Uncertainty Factors for Mixture Risk Assessment." *Archives of Toxicology*, October 8, 2018. DOI 10.1007/s00204-018-2325-6.

Wiecek, Witold, Jean-Lou Dorne, Nadia Quignot, Camille Bechaux, and Billy Amzal. "A Generic Bayesian Hierarchical Model for the Meta-Analysis of Human Population Variability in Kinetics and Its Applications in Chemical Risk Assessment." *Computational Toxicology*, August 2019, DOI 10.1016/j.comtox.2019.100106.

Rajaram, Sankarasubramanian, Witold Wiecek, Richard Lawson, Betina T. Blak, Yanli Zhao, Judith Hackett, Robert Brody, Vishal Patel, and Billy Amzal. "Impact of Increased Influenza Vaccination in 2-3-Year-Old Children on Disease Burden within the General Population: A Bayesian Model-Based Approach." *PloS One*, no. 12 (2017). DOI 10.1371/journal.pone.0186739.

Software

For Chapter 3 computer code for meta-analyses is included directly in the thesis. For Chapter 2, two computer programs for inference on Bayesian networks were developed:

graph_sampler, version 3.0, C software available at `www.nongnu.org/graphsampler`
rgraphsampler, R package available at `github.com/wwiecek/rgraphsampler`

Contents

Chapter 1

Introduction

The popularity of Bayesian approaches to statistical inference has been growing for several decades. Partly thanks to improvements in the computational capacity of personal computers, the use of Bayesian inference is now widely accepted as a viable choice in many standard cases of statistical analysis, primarily through Markov Chain Monte Carlo (MCMC) methods. The emergence of statistical software such as the BUGS family of programs, JAGS and Stan[1] made Bayesian data analysis easier, faster and more reproducible. Meanwhile, the development of more sophisticated approaches such as variational Bayes and approximate Bayesian computation continues to make Bayesian inference feasible in ever more complex models. The recent popularity of Bayesian theory and applications is not limited to statistical inference, with Bayesian theory contributing to new fields of research such as Bayesian improvements to machine learning models, Pearl's causal inference theory or new models of cognition, e.g. of visual perception (see Chipman et al., 2010; Neal, 1996; Pearl, 2009; Clark, 2016).

The main and most widely appreciated reason for taking a Bayesian approach to statistical inference is its ability to account for prior convictions and the fact that Bayesian probability can be understood intuitively. There are also more practical advantages which are known to Bayesian statisticians but may not be fully appreciated by the statistical research community at large:

[1] BUGS was developed at MRC Biostatistics Unit of University of Cambridge from 1989. It uses Gibbs sampling and is most commonly used on Windows, through its WinBUGS version. All BUGS programs are available at *https://www.mrc-bsu.cam.ac.uk/software/bugs/*. Lunn et al. (2009) provide an overview of BUGS development. JAGS was developed by Martyn Plummer and is available at *http://mcmc-jags.sourceforge.net/*. JAGS is similar to BUGS both in syntax and in that it is also a Gibbs sampler (Plummer, 2003). It is however written in C++, which makes it easier to use and modify on different platforms. Stan (*http://mc-stan.org*) uses a new Hamiltonian Monte Carlo approach (see Hoffman and Gelman, 2014, for details) and is also a Turing-complete programming language. It interfaces with popular programming languages and software (R, Python, Julia, MATLAB) and can use user-defined C++ functions.

- Bayesian inference is particularly well-suited to multi-level modelling, *i.e.* such common statistical procedures as mixed effects regression modelling or meta-analysis problems where heterogeneity across units of analysis needs to be correctly captured. MCMC approaches can correctly propagate uncertainty in complex non-linear models with multiple hierarchies and the implementation of such models is relatively easy and intuitive (see Kruschke and Vanpaemel, 2015, for an overview in the context of hierarchical models).

- Outputs of Bayesian data analyses are often distributions of model parameters under some assumed data generating process. This makes model checking and selection simple and intuitive, while also helping to elucidate the underlying modelling assumptions through posterior predictive checking (see Gelman et al., 2013, Chapter 6).

- By "staying Bayesian", researchers can avoid problems with how the null hypothesis significance testing (NHST) is used in statistical research, a topic which has been debated widely within the scientific community: Cohen (1994); Ioannidis (2005) provide seminal examples of problems with NHST as it is applied; Vidgen and Yasseri (2016) provide Bayesian context for these arguments; recently, proposals for addressing the problem were made by Wasserstein and Lazar (2016); McShane et al. (2019); Ioannidis (2018, 2019).

Equally, there are also both theoretical and pragmatic issues (some perceived, some justified) which seemingly stop the Bayesian approaches from being more widely adopted. On the conceptual side, many criticisms focus on the assumptions necessary to conduct Bayesian analysis (often to do with specification of priors) or, more generally, the lack of adequate mathematical and statistical foundations in Bayesian statistics (including, but not limited to, the subjective nature of probability inherent in Bayesian approaches), at least as they are typically applied. A prominent recent example of such criticism is work by Deborah Mayo (e.g. Mayo, 2018). Gelman and Robert (2010) provide an overview of this debate in its historical context and from a Bayesian viewpoint.

However, the main barriers to wider adoption of Bayesian statistics by the research community are arguably more pragmatic. Firstly, despite all the recent improvements, Bayesian approaches are still more computationally intensive and require more technical knowledge to implement and perform. They are particularly daunting to use for non-statisticians when compared with various existing "out of box" statistical software alternatives. Secondly, even if the benefits of taking a Bayesian approach (e.g. use of informative priors, propagation of uncertainty) are appreciated in theory, it may not be obvious to practitioners how they can be of practical benefit in the context of their research question, often due to lack of existing examples of such work in their field of interest.

It is desirable, then, to extend Bayesian inference models specifically to fields of research where they are less popular, as well as to introduce replicable models

and tools which can make Bayesian analysis easier to perform. This is the overall aim motivating this thesis. In particular, we aim here to present uses of Bayesian inference in three areas of epidemiological and biological modelling: 1) modelling of networks, especially including cyclical structures (which often occur in biological systems), 2) characterising population variability for toxicological risk assessment through meta-analysis, and 3) infectious disease modelling of influenza. In all three cases, Bayesian methods are the most appropriate to answer the research question (for reasons which we outlined above) but are not yet widely adopted due to their relative computational and methodological complexity.

Each research project included in this thesis had a different motivation and genesis. For the part of the thesis dealing with Bayesian networks, the problem to solve was theoretical, albeit motivated by a realistic research question of how to account for the presence of cyclical structures in networks. While the proposed solution is not specific to any type of network, the problems and examples behind the research came from modelling biological networks, where cyclical structures are both common and follow regular patterns. The work on this problem was conducted in collaboration with Frédéric Bois (INERIS, Certara UK) and Ghislaine Gayraud (UTC Compiegne).

For the meta-analyses of toxicokinetic data, the models presented here were initially used to inform default values of parameters in generic toxicokinetic and toxicodynamic models in humans, as part of a project sponsored by the European Food Safety Authority.[2] Following the completion of the project, the meta-analysis models were then developed further by the author to provide a generic and replicable approach which can be used in the risk assessment of chemicals.

The work on modelling influenza epidemics was a part of a multi-year, retrospective study conducted by Analytica Laser to understand the public health impact of changing influenza vaccination policy in the United Kingdom. The methodological work was therefore motivated primarily by the challenge of developing a complex Bayesian model that would be fit for this purpose and could work with the wide variety of data inputs which were obtained for the study.

In what follows, we provide a short overview of the three areas of research and outline the problems which were solved using Bayesian approaches.

[2] *Integrating toxicokinetics in chemical risk assessment: application to human, animal and environmental risk assessment*, tender OC/EFSA/SCER/2014/06. Objective of the project was "to develop a suite of tools and models that integrates toxicokinetic (TK) data and can be used to predict TK parameters for human risk assessment of single and multiple compounds", which in turn required literature review to describe means and variability of physiological, biological and toxicokinetic parameters in humans, to be used in the generic tools.

1.1 Learning Bayesian networks

A graphical model is a graph where vertices represent random variables and edges represent cause-effect relationships between the variables. Formal use of graphical models in statistics dates back to causal path analysis methods by Wright (1921).[3] At their basic level, the graphical models are tools for describing conditional independencies of random variables, but they are often used with the implicit goal of making statements about causal relationships between variables. Arguably, the most popular and methodologically consistent framework for working with graphical models is due to Pearl (2009), who posited a theory of causal inference based on Bayesian networks.[4]

A Bayesian network (BN) is a pair consisting of a graphical model, where the graph $\mathcal{G}$ (of n nodes) is directed and acyclic (DAG), and its associated probability distribution $p = (x_1, \ldots, x_n)$ on the random variables associated with nodes. In a BN, p admits chain rule factorisation $p = \prod p(x_i|Pa_i)$, where Pa_i are parents of node i. In other words, x_i is independent of its non-*descendants*, conditional on its *parents*.[5]. Under Pearl's formalism, Bayesian networks carry a causal interpretation and the objective of inference is usually learning from data which random variables are causally linked. Other objectives may be learning parameters (when dependence between variables is given by a parametric model) or even simultaneously learning parameters and structure.

Bayesian networks are particularly popular in learning biological networks, as a part of systems biology approach, part of which focuses on collection and modelling of large-scale quantitative data. Relationships in biological networks are often non-linear and involve long cause-effect paths. Conversely, both large datasets and contextual information on network structure are often available. All these factors contribute to the popularity of algorithmic approaches to studying biological networks. Networks that are object of such studies may be cell signalling pathways, metabolic pathways or gene regulation networks (Sachs et al., 2005; Husmeier, 2003; Blair et al., 2012; Friedman et al., 2000; Bansal et al., 2007; Werhli and Husmeier, 2007).

Despite their name, which was chosen to emphasise "reliance on Bayes's conditioning as the basis for updating information" (Pearl, 2009, pg 14), the problem of learning structure of a Bayesian network does not typically require the use of Bayesian methods for inference. The approaches to learning structure are typically

[3] Significantly, the first uses of these models by Wright were in genetics.

[4] Other notable examples of frameworks for causal inference which stemmed from Wright's work on path dependency include structural equation modelling, first popularised in econometrics and then psychometrics (Haavelmo, 1943); Donald Rubin's potential outcomes framework (Rubin, 2005); Robins' causal inference models in epidemiology (e.g. Robins et al., 2000).

[5] In a directed graph, parents of a node X_i are nodes with an edge leading into X_i. Set of all parents of X is denoted as Pa_i. The children of X_i are nodes to which there is an edge from X_i. Descendants of X_i are all nodes to which there is a directed path from X_i.

divided into two categories: conditional independence testing methods and score-based methods. In the former, the focus is on finding sets of graphs which imply the same conditional independencies (see, for example, Spirtes et al., 1993; Peters et al., 2013). In score-based methods, each DAG $\mathcal{G}$ gets assigned a score $s(\mathcal{G}|\mathcal{D})$, conditional on data $\mathcal{D}$. Typically the approach is then to algorithmically search for the highest scoring graph (Chickering, 2002). Notably, one typical choice for the score s takes a Bayesian approach by setting

$$s(\mathcal{G}|\mathcal{D}) = \log(p(\mathcal{D}|\mathcal{G})) + \log(p(\mathcal{G})), \tag{1.1}$$

where $p(\mathcal{G})$ is the prior distribution over DAGs and $p(\mathcal{D}|\mathcal{G})$ is the likelihood, with model parameters integrated out (Geiger and Heckerman, 2002).

Learning Bayesian networks involving motifs

As mentioned earlier, biological networks often contain regular structures, such as patterns present in gene regulation networks (GRN). An overview of such patterns ("motifs") is provided by Alon (2007). These patterns can be cyclical, forming e.g. feed-back loops. The knowledge that certain patterns are more likely can be exploited by using a Bayesian score (1.1) and putting an appropriate prior $p(\mathcal{G})$ on the occurence of these patterns.

Previously, Bois and Gayraud (2015) have shown how such priors can be used to synthetically create networks containing motifs by using MCMC methods, with GRN of *Eschericha coli* used as an example. Priors on motifs have since been included as a part of *graph_sampler* software for MCMC inference on Bayesian networks (see Datta et al., 2017, for overview). In contrast to many existing methods for inference on BN structure, *graph_sampler* is "fully Bayesian" in a sense that the output of the algorithm is a probability distribution over space of graphs, while in other methods typically only the best graph (or Markov equivalence class of such graph) is returned.

However, by definition, the Bayesian network approach assumes acyclicity and therefore the inferred graphs do not permit cycles. This problem has previously been addressed in literature in many different ways, starting with work by Spirtes (1995); Richardson (1996); Koster (1996); commonly the focus is on studying properties of a wider class of graphs (for example, in the case of Koster, "reciprocal" graphs) and devising inference algorithms within that class.

In this work we propose an alternative approach which takes advantage of the existing MCMC algorithm. In this approach we retrieve properties of Bayesian networks not directly on the scored graphs (DAGs) but on *condensed* graphs, which treat cycles as single nodes and by definition are acyclic. Algorithmically, condensing a

graph depends on identifying its *strongly connected components*[6], which can be done in linear time. The MCMC algorithm will then be modified to take samples from condensed graphs and score them; in this work a linear Gaussian model will be used, but the overall approach to sampling from the space of condensed graphs that we will present does not depend on the statistical model.

1.2 Bayesian meta-analysis for risk assessment

Our second research project focuses on synthesising evidence on risks of exposure to chemicals. The meta-analysis approach which we will discuss can be seen as a special case of hierarchical linear models, which are common in Bayesian inference, including risk assessment. Let us introduce a simple hierarchical model to serve as a guiding example for this section. Assume that k randomised experiments have been carried out, each measuring the same continuous outcome y_{ij}, with $i = 1, \ldots, k$ studies, with study i reporting $j = 1, \ldots, n_i$ observations. The likelihood model is given by

$$\begin{aligned} y_{ij} &\sim \mathcal{N}(\tau_i, \sigma_y^2), \\ \tau_i &\sim \mathcal{N}(\tau, \sigma_\tau^2). \end{aligned} \tag{1.2}$$

The three *hyperparameters* of this model are the grand mean τ and two scale parameters σ_y, σ_τ. When only aggregate (summary) data are known, *i.e.* means $y_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ together with their standard errors se_i, as is common when analysing data from reviewed literature, the model can be adapted easily to yield a meta-analysis model. [7] Typically the main consideration for a researcher is made with regards to σ_τ, as it determines how the estimates of mean y_i's will be "pooled" together (or not), with the two extremes being $\sigma_\tau = 0$, the "full pooling" ($y_1 = \ldots = y_k$) and $\sigma_\tau = \infty$ being "no pooling".[8] More plausibly, there is some heterogeneity across studies, but their effect is still related, i.e. $0 < \sigma_\tau < \infty$. In that case we have to infer σ_τ from a balance of data and the prior, choice of which is discussed in detail by Gelman (2006).

[6] As we will discuss later, in graph theory not all SCCs are cycles, therefore when discussing theoretical properties we refer to components, rather than cycles.

[7] Classic example of which is the "eight schools" educational intervention experiment described first by Rubin (1981) and used in various tutorials and as examples in Bayesian inference software. If study-specific covariates are introduced into the model for y_i's, we call such type of meta-analysis a *meta-regression* model, sometimes specifically to contrast it to models with no covariates. The model presented in this thesis is a meta-regression model, but we prefer to refer to it more generically, as a meta-analysis model.

[8] We can also say that pooling refers to *fixed* or *random* effect models in a sense that these terms are typically referred to in statistical modelling. Gelman (2005, Section 6) discusses how these terms relate to multilevel modelling and points out multiple possible definitions and interpretations of terms "fixed effect" and "random effect".

When the real pooling is partial, taking the Bayesian approach offers an additional advantage: the result of inference is not only a posterior distribution for the mean effect τ, but also a predictive distribution for the mean of a "new" (or, rather, a "hypothetical") study, τ_{k+1}. This distribution can be used for model checking and cross-validation, but sometimes it is also of interest in and of itself. In a meta-analysis model for risk assessment, which we will present later, multiple hierarchies are included: study, compound, subgroup. When criticising such model and deriving predictions, the Bayesian approach to meta-analysis allows us to derive predictive distributions at various levels of hierarchy, e.g. predicting new studies of observed compounds or means for hypothetical compounds, while simultaneously accounting for uncertainty in other parameters.

Hierarchical Bayesian meta-analysis models are used in many fields, albeit often requiring quite complex modifications. Choice examples that illustrate the variety of approaches are: non-linear meta-analysis models in pharmacology measuring response to drug administration (Weber et al., 2018), meta-analysis models of hierarchical quantiles of a continuous outcome applied to economic data (Meager, 2019), and network meta-analysis and meta-regression models of survival data (of time to outcome) to evaluate health outcomes (Jansen, 2011). These three examples all use different models and address different types of research questions, but all three can be seen as domain-specific adaptations of the same hierarchical modelling idea, where the central question is that of characterising the effect of some intervention while taking into account differences in experimental conditions and/or heterogeneity of studied populations.

Risk assessment and toxicokinetics

Characterising the relationship between exposure to chemicals and the toxic response is essential to chemical risk assessment in humans, which aims to establish safe levels of exposure to chemicals. In practice, the recommended value is established in two steps: first by determining the reference value, the dose associated with adverse event[9] and then dividing it by an *uncertainty factor* (UF), a value that accounts for uncertainty in available data. Interspecies and intra-human variability were traditionally addressed by applying a default UF of 100 to the reference value in test species. Many refinements to this approach have been proposed, in particular starting with Renwick and Lazarus (1998), who proposed to subdivide uncertainty factors between toxicokinetics (TK) and toxicodynamics (TD) and inform them in a chemical-specific manner.[10]

[9]This reference dose can be determined using no-observed-adverse-effect-level (NOAEL), lowest-observed-adverse-effect level (LOAEL) or, more recently, benchmark dose (BMD) methods, which use dose-response modelling (see Hardy et al., 2017b).

[10]Since the uncertainty in our context is related to variability, the term "variability factor" (or "uncertainty and variability factor") has been proposed, but for simplicity we will refer to all quantities that are used to divide the reference value as UFs.

Meta-analysis offers a way to synthesise information on human variability in TK/TD and help refine UFs as a part of the "weight of evidence" (WoE) approach to risk assessment. WoE is defined by World Health Organisation as "a process in which all of the evidence considered relevant for a risk assessment is evaluated and weighted". While both qualitative and quantitative approaches are used in practice, the development of new tools and sources of data has in recent years led to a shift to more quantitative approaches. European Food Safety Authority's recent guidance document for WoE (Hardy et al., 2017a) includes meta-analysis as first on the list of quantitative approaches.[11]

Meta-analysis for risk assessment

Our objective is to develop meta-analysis models for the purpose of informing risk assessment of chemicals. Similarly to the examples of meta-analysis models given above, we are required to modify the simple model given by (1.2) in three ways: to suit available data, our modelling assumptions and the objective of analysis. Briefly:

- *Data.* A lot of toxicological data comes from studies with very small sample sizes, including volunteer studies. For example, a literature review on which a later example will be based (and which can be seen as representative) reported a quarter of studies with sample size of 2 or 3 and three quarters with samples of 10 or less. In addition, many different measures of dispersions are used in literature, with individual-level data sometimes available.

- *Model.* While the reported values of quantities of interest (e.g. parameters describing the speed at which a compound is metabolised) are highly variable, these values are often related in a way that can be informed by prior information (e.g. ratios of such parameters between different phenotypic groups of population depend on drug-specific properties which are known from other studies). Thus some additional constraints can be introduced in the model to account for our prior knowledge of this relationship.

- *Objective.* Assessing risk of exposure to chemicals is concerned with the impact of exposure not on the population means but on the entire distribution of quantities of interest. Therefore, the output of the meta-analysis model should characterise either the entire distribution or specific tail area probabilities in at-risk groups (e.g. the 1% of slowest metabolisers in a specific ethnic subgroup of the general population).

[11] In particular, the guidance document notes that "statistical models may also be used for meta-regression to explain the variability between studies as a function of explanatory variables, for example, population characteristics or study quality issues. They are able to describe uncertainties through confidence intervals and probability distributions. Other types of statistical methods (e.g. Bayesian methods) are also useful for synthesising multiple sources of evidence."

As the model is motivated by a practical problem, we will present both the generic modelling considerations and adaptations of the model to specific cases: polymorphic enzymes in single chemicals and in case of mixtures of chemicals.

1.3 Bayesian inference in modelling influenza

Modelling of infectious diseases in humans is an important area of mathematical modelling. A commonly used framework divides the host population into subgroups, identified according to their risk status (e.g. age) and disease status or "state". As an example, consider a 4-compartment model where the population is divided into four states: susceptible, exposed, infected and recovered (SEIR)[12]. Such a model is appropriate for diseases where there is a delay between acquiring the virus and being infectious, such as influenza (a latent period). Under a SEIR model, dynamics of the disease over time are determined by ordinary differential equations associated with each compartment:

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta SI}{N} \\ \frac{dE}{dt} &= \frac{\beta SI}{N} - \sigma E \\ \frac{dI}{dt} &= \sigma E - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{1.3}$$

It is assumed that each physical or social contact between infectious and susceptible individuals can result in a new infection, with some fixed probability; this assumption is captured by parameter β, a factor of infection probability and the number of contacts made. Average lengths of exposed and infectious periods are given by $1/\gamma$ and $1/\sigma$ respectively.[13]

While compartmental models are not able of capturing all of the dynamics that are relevant to spread and course of infectious disease (especially reflecting individual's contact pattern), the basic SEIR model can be modified in many different ways

[12]Modelling of spread of infectious disease in populations can be divided into two classes: populational (compartmental) and agent-based. Gilbert (2007) provides a cross-disciplinary introduction to agent-based models. An influenza-specific model operating at level of individual agents is presented by Chao et al. (2010). Here, only the compartmental models are considered.

[13]Under a model with one exposed (infectious) compartment, the length of time spent in these compartments is given by exponential distribution. Wearing et al. (2005) show how a more appropriate model in the case of influenza is one with multiple (repeated) compartments, giving a gamma-distributed latent (infectious) periods. That is the approach we will take in this work.

to describe epidemics in a more detailed manner, taking into account *e.g.* immunisation, treatment, changing contact patterns. Through a combination of statistical inference and simulaton, SEIR models are commonly used to address public health policy questions in seasonal influenza; see, for example, work by Gani et al. (2005) on impact of antiviral medication, Vardavas et al. (2007) on increasing voluntary vaccinations, Kamal et al. (2017) to evaluate health economics of influenza treatment. The work we will present here follows these SEIR approaches, but focusses primarily on inferring the real epidemics to inform public health policy.

Seasonal influenza and public health

Seasonal influenza is a big public health risk. While in most people the course of the disease is not severe, influenza can lead to death, especially in the elderly and high-risk individuals, due to complications which can develop after the infection. From a public health perspective, the burden of influenza[14] also manifests in lost days of work, healthcare utilisation (doctor visits), and hospitalisations.

For these reasons, studying the impact of seasonal influenza vaccination or of antiviral treatments for influenza is an important part of informing the public health policy in all developed countries. The dynamics of infection have direct implications for such studies, since the individuals who are likely to transmit the disease are, on average, not the individuals most at risk for hospitalisation or death. Studies of social contact patterns reveal that children and young adults appear to be primary "vectors" for the seasonal diseases, because of their high average number of daily contacts, including physical contacts (Mossong et al., 2008). Meanwhile, the two groups most at-risk are infants and elderly, due to their increased mortality and hospitalisation risk following an infection.

The problem of choosing the optimal public health policy is complicated by influenza surveillance. While influenza testing and seasonal epidemic monitoring are common in developed countries,[15] precisely modelling the true number of cases of influenza is difficult, as the spread of influenza depends on factors relating to biology, social behaviour, demographics, healthcare system and more. Even given a correct model (*i.e.* one that is capable of approximating the real number of influenza cases well enough in the context of the research question), relating the latent number of cases to observed quantities has two problems. First, the propensity of infected individuals to seek medical care is uncertain and, crucially for understanding the

[14] "Burden of disease" being a generic term for quantifying the impact of disease in terms of health events and its economic impact.

[15] The Centre for Disease Control in the United States publishes weekly reports that include data on virus testing, general practitioner (GP) consultations, hospitalisations, and deaths attributable to influenza. Such data can additionally be broken down by age, risk, location. Similarly detailed reports are available on the Internet in the United Kingdom, through Public Health England and for European Union countries through European Centre for Disease Control. These sources also comprehensively report vaccination rates and presence of other seasonal viruses over time.

vaccination policy, varies with age. Second, patients with influenza-like symptoms may be affected not by influenza but by another seasonal virus. Third, , all of which need to be accounted for in the model and linked to the observed quantities.

Inferring disease burden and influenza epidemics

Because of the difficulties in estimating the underlying epidemic (numbers of influenza infections over time, across groups of interest), the burden of influenza is often estimated by statistical analyses of the observable quantities alone, using, for example, regression approaches. Particularly common are excess risk approaches (see, for example, Simonsen et al., 1998, 2005), which focus on analysing seasonal trends in *e.g.* deaths to decide what proportion of them can be attributed to influenza epidemics. However, since these methods do not infer the data generating process (influenza epidemic), they are less suited to directly address questions about the impact of hypothetical public health policies on the burden of influenza, unlike e.g. SEIR models which can be used to generate counterfactual or compare a number of simulated scenarios, as in the examples we cited above.

When inferring the parameters of influenza epidemics and surveillance model, there is a strong rationale for using a Bayesian hierarchical approach. First, they offer a method for joint estimation of quantities generated by the non-linear disease dynamics, allowing us to join many data sources under the same model. Second, since parameters of epidemic and surveillance model are highly uncertain, it is important that the chosen method allows for the propagation of uncertainty. Third, in this setting informative priors can be used on almost all of the model parameters.

Despite this, Bayesian inference in infectious disease models of influenza is not common. Notable examples of recent modelling include work Baguelin et al. (2013) (also on vaccination) and its recent extension by van Leeuwen et al. (2017) or the Bayesian inference and forecasting approach to influenza epidemics by Osthus et al. (2017). We speculate that the main reason for this lack of popularity is the difficulty in implementing MCMC methods for inference on compartmental models, which is a computationally and statistically complex task. This is exemplified by the fact, that the papers we cited above are largely concerned with methodological, computational issues, in contrast to the simulation or regression approaches we referenced.

Measuring impact of vaccination

The work which presented in this thesis uses SEIR models to investigate the impact of introducing routine influenza vaccination for 2 and 3 year olds in England. Such policy was introduced by Public Health England starting in 2014 and at the time of writing the 2-3 year olds in England are still being offered routine vaccinations

(NHS, 2017). The objective of the study on which our work is based was to make estimates specific to different influenza seasons, specifically from 2010 to 2014, i.e. prior to change of the policy and in the first season where it was implemented.

This approach required combining statistical inference on parameters with simulations to quantify impact for each season, as described above. In the first three seasons, the counterfactual scenario was the routine vaccination of 2 and 3 year olds, while for the 2013-2014 season the situation was reversed and the counterfactual assumed no vaccination program. Data collection covered health care records for 7% of the population of the UK and various other sources on GP consultations, hospitalisations, deaths, influenza testing, presence of another seasonal virus. A separate retrospective study was devoted to collection and reporting of data. Thus one of the main challenges of this research was in devising a model that could simultaneously account for all of the available data while providing the benefits of the Bayesian approach which we listed above.

Structure of this thesis

In the three chapters that follow we show how the three problems outlined above have been solved. The new approach to modelling cycles in Bayesian networks and its implementation are presented in Chapter 2. Chapter 3 provides a Bayesian model for meta-analysis of toxicological data together with case studies. Chapter 4 describes the Bayesian model of influenza vaccination policy. Each chapter starts with references to published (or submitted) articles and a short overview of the author's contributions. The work is summarised and further directions of research are discussed in Chapter 5.

Chapter 2

Learning Bayesian networks

The work presented in this chapter comprises a publication and statistical inference software. Wiecek et al. (2019a) (submitted) describes the methodological contribution of this work, where the focus was on devising a model to work with cyclical structures. The described approach has been efficiently implemented in *graph_sampler*, an open-source software available at *www.nongnu.org/graphsampler* through modification of MCMC sampler. In addition, an R package was developed, which includes all of *graph_sampler*'s functionality (not limited to inference on cycles). The software contribution is presented at the end of this chapter.

Structure learning of Bayesian networks involving cyclic structures

Witold Wiecek*, Frédéric Y. Bois†, Ghislaine Gayraud‡

Abstract

Many biological networks include cyclic structures. In such cases, Bayesian networks (BNs), which must be acyclic, are not sound models for structure learning. Dynamic BNs can be used but require relatively large time series data. We discuss an alternative model that embeds cyclic structures within acyclic BNs, allowing us to still use the factorization property and informative priors on network structure. We present an implementation in the linear Gaussian case, where cyclic structures are treated as multivariate nodes. We use a Markov Chain Monte Carlo algorithm for inference, allowing us to work with posterior distribution on the space of graphs.

1 INTRODUCTION

Large-scale gene expression studies have invigorated interest in exploratory methods for evaluating patterns of association between random variables. The large number of random variables potentially considered and relatively small data sets challenge known structure learning approaches both conceptually and computationally. Graphical models are often used for to represent network structureand for statistical inference (Lauritzen, 1996). They depict the random variables of interest as nodes in a graph, and conditional independence statements about them by presence or absence of graph edges. We focus on Bayesian networks (BNs) which represent probability distributions by means of directed acyclic graphs (DAGs) and are popular in learning structure of biological networks (e.g. Husmeier, 2004; Husmeier and Werhli, 2007). Under additional assumptions, described for example by Pearl (2009, Chapter 1), directed edges of BNs can correspond to causal relationships between nodes. In BN models the joint probability distribution can be factorized between nodes and evaluated easily.

Yet, there are important cases in biology and other domains of study where we do want to consider cyclic structures such as feedback loops, common in gene transcription regulation networks and other biological networks (Alon, 2007). In such cases

*Certara UK Ltd, 5th Floor Front, Audrey House, 16-20 Ely Place, London EC1N 6SN, United Kingdom

†Certara, Simcyp division, Level 2-Acero, 1 Concourse Way,Sheffield S1 2BJ, United Kingdom

‡Sorbonne Universités, LMAC, Université de Technologie de Compiègne, France

DAGs cannot be used directly, as they do not offer a sound representation of an essential feature of the networks under analysis. Dynamic Bayesian networks (DBNs) offer an alternative, by unrolling cycles in time. However, they multiply the number of nodes by the number of observation times and require dense and extensive data series, as discussed by Ghahramani (1998).

We present a different approach to modelling cyclic structures within a graph,. We contract such structures within the graph to derive an associated acyclic graph. The contracted structures are treated as multidimensional random variables. For inference on graph structure, we use a score-based implementation in the linear Gaussian case. Our approach is fully Bayesian, with scores being Bayesian prior predictive densities. Our procedure uses the factorization property of BNs and, to our knowledge, is novel. We implemented it in *Graph_sampler*, an efficient C language software for simulated network generation and Bayesian inference on network structures.

The informative priors we use, including on cyclical structures, imply that scores between two graphs may differ even if those graphs entail the same conditional independencies. In this sense our approach is an extension to previously proposed Bayesian approaches to network inference, such as the approach by Mukherjee and Speed (2008), where the focus is on working with the full posterior distribution.

This paper is organized in two parts. First, Section 2 presents the statistical model, broken down between graph theory background, graph priors (including hyperparameters) and derivation of likelihood (graph score). Then, Section 3 presents choice of hyperparameters and examples of applications in graphs which involve cyclic structures, including computational benefit to Markov Chain Monte Carlo algorithms.

2 STATISTICAL MODEL

Methods for learning structure of a Bayesian network (presence or absence of edges between fixed nodes) can be categorized as either test-based methods for conditional independence or score-based methods. The latter tend to give more accurate results, according to Acid et al. (2004) and Cooper and Herskovits (1992), but their major disadvantage is the computational cost: since the number of possible graphs to consider grows superexponentially with their number of nodes, exact inference on structure is a hard problem.

In what follows we use a score-based method in a Bayesian framework. For any directed graph $\mathcal{G}$ (not necessarily acyclic) we define graph's score $s(\mathcal{G}|D)$ conditionally on observed data D. It is proportional to weight of evidence $s(D|\mathcal{G})$ and prior distribution over space of graphs $p(\mathcal{G})$. We define score for any directed graph, not necessarily acyclic.

Term $s(D|\mathcal{G})$ is obtained from the graph's prior predictive density (p.p.d.), that is the data likelihood integrated over all of the model parameters. Such an approach is computationally more efficient than calculating the full posterior function, and parameter values are not needed to make inferences about structure. However, for brevity we still refer to this quantity as posterior, even though some parameters have been averaged out. In the data model, we make use of prior conjugacy, which helps quickly evaluate the p.p.d. A Metropolis-Hastings Markov Chain Monte Carlo (MCMC) algorithm can then be used to sample random graphs from their posterior distribution, *cf.* Yu et al. (2004), Zhou et al. (2004), Datta et al. (2017).

As we will show later in this Section, graphs which imply same conditional independencies may not have the same score. For acyclic graphs, this is due to use of informative priors. When cyclic structures are allowed, this inequality may also arise by choice of hyperpriors which can promote or penalise the occurence of cycles. Before proceeding with a description of the statistical model, we introduce graph theory definitions on which the statistical model depends. The rest of this section describes the priors and data likelihood we use.

2.1 Graph model

In what follows, we assume that $\mathcal{G} = (V, A)$ is a directed graph (with vertex set V and set of directed edges A), of a given order $N = |V|$. We do not require for $\mathcal{G}$ to be acyclic, but edges from a node to itself (*autocycles*) are not allowed. We use terms *graph* and *network* interchangeably.

We say that a graph is *strongly connected* if for every pair of vertices there exist paths in each direction between the two. A *strongly connected component* (SCC) of a graph is a maximal subgraph that is strongly connected. By definition, every cycle is a strongly connected (although not maximal) subgraph. Not all SCCs are cycles, however; *e.g.* a "flat eight" graph of three nodes $A \to B \to C \to B \to A$ is strongly connected but B is traversed twice to get from A to C, hence it is not a cycle. We call single node components *ordinary*. For each graph we can create a partition of its vertices into sets of strongly connected components. We denote such partition by $\mathrm{SCC}(\mathcal{G})$. It can be performed in linear time, as first proposed by Tarjan (1972). Since most of the strongly connected components which we encounter in structure learning of biological networks are graph cycles, we will also interchangeably use term "cyclic structures" throughout the paper.

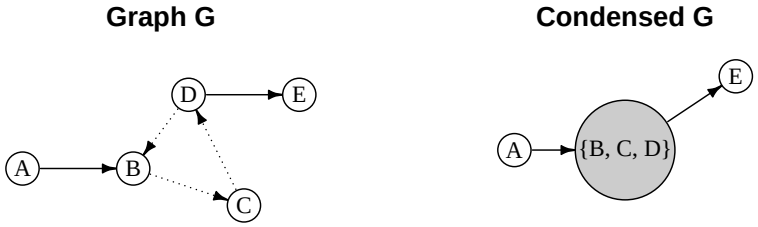


Figure 1: Graph $\mathcal{G}$ with 5 nodes where nodes B, C and D form a cycle and are contracted to a single node after condensing the graph. Here $\mathrm{SCC}(\mathcal{G}) = \{A, \{B, C, D\}, E\}$

For any directed graph $\mathcal{G}$, we can create an associated *condensed graph*, $\mathcal{G}_c$, by repeatedly contracting edges (replacing a pair of vertices connected by an edge by a single vertex, retaining all directed edges) in each strongly connected component until each component corresponds to a single vertex. By construction such graph is acyclic. An illustration of the process is provided in Figure 1.

For DAGs, a *Markov equivalence class* of graphs is a set of DAGs that have the same skeleton (set of edges without regards to direction) and v-structures (sets consisting of a child and its two parents that are not themselves connected), as shown by Verma and Pearl (1992) in the context of causal inference. Members of the equivalence class encode same conditional independencies. Various algorithms have been proposed to learn Markov equivalence classes of causal graphs, *e.g.* by Chickering (2003).

When cyclic structures are present and we are working with condensed graphs, we assume that the conditional independencies implied by the Bayesian network are only the ones that are implied by the condensed graph $\mathcal{G}_C$. As $\mathcal{G}_C$ is a DAG, we can take advantage of the Markov property and factorise the joint probability distribution over nodes of the condensed graph $\mathcal{G}_c$. We then model the initial $\mathcal{G}$ by considering multivariate normal distributions on each of the strongly connected components.

2.2 Priors on graph structure

All possible graphs of a certain size are not equally plausible *a priori* and we should consider that prior knowledge on graph structure in our inferences. Distributions and parameterizations of the following priors which have been previously described by Mukherjee and Speed (2008) and by Datta et al. (2017):

- *Bernoulli priors* on existence of individual (directed) edges, specified by providing a square matrix of edge probabilities. Gene association studies can provide this type of prior knowledge.
- *Concordance prior* between the graph adjacency matrix and an edge requirement matrix, where each edge is classified as desired, not desired or no preference. This penalizes networks too different from a canonical one (although, tuning this pseudo-prior is not very easy).
- *Degree prior* on the distribution of node degrees $d(v)$ in the graph, using a power law with parameter γ. The degree distribution of many physical networks appear to follow approximately such a power law (Barabási and Albert, 1999).
- *Edge count prior* on expected graph size.
- *Motif prior* on count of triangular feed forward and feedback loops in the network, as discussed in Bois and Gayraud (2015).

All the above priors are specified on the input graph $\mathcal{G}$ (and not $\mathcal{G}_c$). To work with cyclic structures, we introduce two additional structure priors on strongly connected components:

Prior on number of strongly connected components. Consider partition of graph $\mathcal{G}$ into $\mathrm{SCC}(\mathcal{G})$, with ordinary nodes discarded. We define a prior on the number of strongly connected components (of at least two nodes) $|SCC(\mathcal{G})|$ via a Poisson distribution: $|\mathrm{SCC}(\mathcal{G})| \sim Poisson(\lambda_{\mathrm{SCC}})$, with $\lambda_{\mathrm{SCC}} > 0$.

Prior on the size of strongly connected components. We also define a prior p_{SCC} on the size of all components present in $\mathcal{G}$, using a power law:

$$p_{\mathrm{SCC}} \propto \prod_{S \in SCC(\mathcal{G})} (|S|)^{-\gamma_{\mathrm{SCC}}},$$

with $\gamma_{\mathrm{SCC}} > 0$.

The global prior probability of graph, $p(\mathcal{G})$, is then proportional as a product of all the priors specified.

2.3 Data likelihood

We now turn to calculating the (integrated) data likelihood, *i.e.* the p.p.d. $s(D|\mathcal{G})$. We want to define that score for any directed graph $\mathcal{G}$, including graphs with cycles. Our aim is to fall back on the key feature of Bayesian networks (the factorisation of its associated probability distribution), and we achieve that by condensing $\mathcal{G}$ into the acyclic $\mathcal{G}_c$.

Let $D = \mathbf{x} = (x_1, ..., x_N)$ denote the observed data on N nodes, where x_i is a n-dimensional vector, with n the number of data points per node. When we condense $\mathcal{G}$, we bifurcate its nodes (and corresponding x's) into nodes obtained by contraction (corresponding to strongly connected components of at least two nodes) and ordinary nodes, for which no contraction was needed (corresponding to strongly connected components of single node). Thus we represent x as:

$$\mathbf{x} = ({x_1}^D, {x_2}^D, ..., {x_{N_1}}^D, \mathbf{x_1}^{\mathbf{L}}, \mathbf{x_2}^{\mathbf{L}}, ..., \mathbf{x_{N_2}}^{\mathbf{L}})$$

where N_1 denotes the number of ordinary nodes and thus various ${x_i}^D$'s are just relabelled x_i's from the original data set. N_2 represents the number of strongly connected components of size at least 2. For $j = 1, 2, ..., N_2$, we denote $\mathbf{x_j}^{\mathbf{L}} = (\mathbf{x_{j_1}}, ..., \mathbf{x_{j_m}})$ a data set of the 'm' node members of the j-th component. Given this partitioning of the graph data, for any given graph G the score can be factorized into a product over ordinary nodes and strongly connected components:

$$s(D|\mathcal{G}) = \prod_{i=1}^{N_1} s({x_i}^D | Pa({x_i}^D)). \prod_{j=1}^{N_2} s(\mathbf{x_j}^{\mathbf{L}} | Pa(\mathbf{x_j}^{\mathbf{L}}))$$

where $Pa(\cdot)$ denotes the parent nodes of "$(\cdot)$" in $\mathcal{G}_c$. When $Pa(x_i) = \emptyset$, $s(x_i|Pa(x_i))$ reduces to $s(x_i)$.

The remainder of this section describes how to obtain the terms $s({x_i}^D|Pa({x_i}^D))$ and $s(\mathbf{x_j}^{\mathbf{L}}|Pa(\mathbf{x_j}^{\mathbf{L}}))$ under a linear Gaussian model.

2.3.1 Integrated likelihood for contracted nodes

Let us consider a strongly connected component of m nodes $j_1, j_2, j_3, \ldots, j_m$. As defined above, $Pa(\mathbf{x_j}^{\mathbf{L}})$ is the set of its parents in the condensed graph $\mathcal{G}_c$, *i.e.* $\cup_{i \in 1,2,...,m} Pa(x_{j_i})$.

We model the distribution of $\mathbf{x_j}^{\mathbf{L}}|Pa(\mathbf{x_j}^{\mathbf{L}})$ using a linear multivariate Gaussian model; hence setting $Y = \mathbf{x_j}^{\mathbf{L}}$, the model can be expressed as

$$Y = X\theta + \epsilon \tag{1}$$

where Y is a matrix of dimension $(n \times m)$, X is the design matrix of size $(n \times k)$ with ones in its first column and $Pa(\mathbf{x_j}^{\mathbf{L}})$ in the remaining columns so that $k =$

$\dim(Pa(\mathbf{x_j^L}))+1$, θ is the matrix involving the coefficient terms of dimension $(k \times m)$, while ϵ is a $n \times m$ dimensional matrix $\epsilon = (\epsilon_1, \epsilon_2, .., \epsilon_n)$ where all ϵ_i's are independent and identically distributed according to multivariate Gaussian distribution $\mathcal{N}_m(0, \Sigma)$. Under this model, the likelihood L is multivariate normal and can be expressed as:

$$L(Y, X, \theta, \Sigma) = \frac{1}{(2\pi)^{nm/2}} |\Sigma|^{-n/2} \exp\left\{ \frac{-tr[\Sigma^{-1}(Y - X\theta)^t (Y - X\theta)]}{2} \right\} \tag{2}$$

where $tr(\cdot)$ denotes the trace of $\cdot$.

For θ and Σ, we consider independent priors, *i.e.*, $p(\theta, \Sigma) = p(\theta)p(\Sigma)$. In order to have an analytically explicit form of the p.p.d. we define an improper (locally uniform) prior on θ, $p(\theta) \propto$ constant. For the prior distribution of Σ, we use an m-dimensional inverse Wishart distribution with hyperparameters (κ, q), denoted $\mathcal{W}_m^{-1}(\kappa, q)$:

$$p(\Sigma) = \frac{|\kappa|^{\frac{q}{2}} |\Sigma|^{\frac{-(q+m+1)}{2}}}{2^{\frac{mq}{2}} \Gamma_m(\frac{q}{2})} \exp\left\{ -\frac{1}{2} tr(\kappa \Sigma^{-1}) \right\},$$

where κ is a positive definite scale matrix and scalar q (degrees of freedom) is strictly positive; Γ_m corresponds to the multivariate gamma function. We refer to this prior as *constant-Gamma*, to distinguish it from other possible models outlined below.

The least square estimate $\hat{\theta}$ for the matrix θ, and the sample co-variance matrix A_0 are given by:

$$\begin{aligned} \hat{\theta} &= (X^t X)^{-1} X^t Y \\ A_0 &= (Y - X\hat{\theta})^t (Y - X\hat{\theta}) \end{aligned} \tag{3}$$

where $\hat{\theta_i}$ is the least square estimate for the $i-th$ column of θ. To define $\hat{\theta}$, we need $(X^t X)^{-1}$ to exist and thus have the constraint $k \leq n$.

The joint posterior distribution of both $(\theta, \Sigma)|(Y, X)$ can therefore be expressed as

$$P(\theta, \Sigma | Y, X) \propto C_0 \cdot P(\theta | \Sigma, Y, X) \cdot P(\Sigma | Y, X) \tag{4}$$

where C_0, $P(\theta|\Sigma, Y, X)$ and $P(\Sigma|Y, X)$ are of the form:

$$C_0 = (2\pi)^{\frac{m(k-n)}{2}} 2^{\frac{m(p-q)}{2}} \frac{\Gamma_m(\frac{p}{2})}{\Gamma_m(\frac{q}{2})} |\kappa|^{\frac{q}{2}} |X^t X|^{-\frac{m}{2}} |\kappa + A_0|^{-\frac{p}{2}}, \tag{5}$$

$$P(\theta|\Sigma, Y, X) = \frac{|X^t X|^{\frac{m}{2}} |\Sigma|^{-k/2}}{(2\pi)^{\frac{mk}{2}}} \exp\left\{ \frac{-1}{2} tr \Sigma^{-1} (\theta - \hat{\theta})^t X^t X (\theta - \hat{\theta}) \right\},$$

$$P(\Sigma|Y, X) = \frac{|\Sigma|^{-(p+m+1)/2}}{2^{\frac{mp}{2}} \Gamma_m(\frac{p}{2})} \exp\left\{ -\frac{1}{2} tr \Sigma^{-1} (\kappa + A_0) \right\} |\kappa + A_0|^{\frac{p}{2}},$$

where $p = q + n - k$.

When θ and Σ are marginalized out from (4), we obtain(5), which corresponds to the prior predictive distribution of the strongly connected component under the *constant-Gamma* model, *i.e.*,

$$C_0 = s(\mathbf{x_j}^{\mathbf{L}} | Pa(\mathbf{x_j}^{\mathbf{L}})).$$

2.3.2 Integrated likelihood for ordinary nodes

In the case of an uncondensed acyclic graph, possible forms of integrated likelihood $s({x_i}^D | Pa(x_i^D))$ have been discussed previously by Datta et al. (2017). With a univariate linear regression model on x_i's parents, using a classical Normal-Gamma conjugate prior (inverse Gamma on the scale and conditional normal on the mean), integrating out these parameters leads to a multivariate Student's t distribution. Zellner and Dirichlet integrated likelihoods are other possible choices and also described therein, together with the choice of likelihood parameters' hyperpriors.

In our case where cyclic structures are allowed, we treat the ordinary nodes as 1-dimensional special cases of the constant-Gamma prior, owing to the fact that the inverse Wishart distribution with parameters q, κ is the multivariate version of the inverse Gamma distribution with parameters $(q/2, \kappa/2)$. We show the equality of the two prior predictive distributions in Supplement S1.

3 MODEL PROPERTIES AND APPLICATIONS

In this section we will discuss three topics: how choice of hyperparameters impacts graph score in SCC cases; how inference on structure is accomplished with MCMC algorithm; some examples of applications of our approach. The examples will illustrate role that priors and SCCs play in both learning network structure and the computational aspect of inference.

3.1 Likelihood equivalence in constant-Gamma case

Geiger and Heckerman (2002) discuss conditions under which graphs in an equivalence class will have the same likelihood. A normal model with inverse Wishart prior is such a case. This notion of equivalence can be extended to *marginalized* likelihoods. Heckerman et al. (1995) present an additional assumption sufficient for marginal likelihood equivalence. It requires that the Jacobian of the one-to-one mapping between two parameters sets associated with two distribution equivalent graphs exists and the priors of the two parameters sets must be equal after applying the change of variables formula.

However, in our case this property no longer holds since we consider independent prior on each parameter set attached to a single term in the likelihood factorization. For the case of a two-node graph, we derive the p.p.d. explicitly in the Supplement S2 and show how $s(D|A \to B) \neq s(D|B \to A)$. In the case of Markov-equivalent DAGs the differences in graph score are due to sampling variance and tend to 0 with increasing n. The score of SCCs will differ from equivalent DAGs due to marginalisation of the likelihood and the difference doesn't tend to 0 with growing sample, but rather depends on the choice of hyperparameters of the inverse Wishart prior distribution of Σ.

Hyperparameters needed to evaluate the graph score under the constant-Gamma model are the scale matrix κ and degree of freedom q. We use inverse Wishart prior as it is conjugate and allows us to obtain the prior predictive distribution analytically. Additionally, this prior, when informative, can be interpreted in terms of equivalent sample size. If $X \sim \mathcal{W}^{-1}(\kappa, q)$ then

$$\begin{aligned} \mathrm{E}(X_{ij}) &= \frac{\kappa_{ij}}{q-m-1}, \quad \text{when } q > m+1, \\ \mathrm{Var}(X_{ij}) &= \frac{(q-m+1)\kappa_{ij}^2 + (q-m-1)\kappa_{ii}\kappa_{jj}}{(q-m)(q-m-1)^2(q-m-3)}, \quad \text{when } q > m+3. \end{aligned} \tag{6}$$

However, there are known issues with inverse Wishart prior: it implies relationships between variances and covariances and uses a single parameter (q) to describe precision on all parameters. When $q > 1$ the prior may be biased when the true variance is low, even with large sample sizes, as discussed by Gelman (2006); see also Alvarez et al. (2014) for a simulation study.

In inference on variance-covariance matrices, it is typical to assume $\kappa = I_m$ (identity matrix of order m) and $q = m + 1$. In such case prior marginal distributions of correlations are uniform on $(-1, 1)$. However, in case of covariance in SCC data this goes against our intuition: typically *a priori* we assume that nodes of an SCCs are going to be strongly correlated; exactly how strongly depends on context and objectives of analysis. By default we propose to set $q = m+1$ and κ to 1 on diagonal elements and to 0.5 on off-diagonal. This creates a monotonic prior on correlation and ensures higher score for SCC than all DAG graphs when the true correlation is higher than 85%-90%. This choice is explored and explained below.

Let us define d as the difference in log scores between a m-node graph where all nodes form an SCC ($\mathcal{G}_{\mathrm{SCC}}$) and a DAG with no conditional independencies ($\mathcal{G}_{\mathrm{DAG}}$):

$$d = \log s(\mathcal{G}_{\mathrm{SCC}}|D) - \log s(\mathcal{G}_{\mathrm{DAG}}|D),$$

conditional on the same data D. Positive d's indicate that SCC scores higher than the DAG.

We now briefly explore the behaviour of d in SCCs of different sizes by means of simulated data and show that it is predictable in ways that may be useful in practical applications. For all the examples presented in this section we generate $n = 10,000$ draws of data of m nodes from multivariate normal distribution with means 0 with each node having fixed variance σ^2 and same correlation ρ with all other nodes. We start with $\sigma^2 = 1$ and vary ρ between 0 and 1. We compare the SCC against one DAG only as at large n all DAGs in the equivalence class will have almost identical scores.

As mentioned, it is typical to set $q = m+1$ and $\kappa = I_m$. Since q is responsible for the precision of the prior and can be interpreted in terms of sample size equivalence, setting a low q is a good default choice. However, even with the default value $q = m+1$, κ has an impact on choice between SCCs and DAGs. Setting $\kappa = I_m$ in this situation leads to uniform priors on correlation. This can be desirable in inference on unknown variance-covariance matrix, but our goal in this case is to distinguish cyclic structures from DAG parts of the network: therefore a sensible choice would be to have a prior that "favours" loops when correlation is higher. We illustrate behaviour of d as a

function of the off-diagonal elements of κ in two panels of Figure 2. With $\kappa = I_m$ the sign of d is not consistent, but when the off-diagonal elements of κ are set to 0.5 everywhere, d is positive when the true correlation in data exceeds 85%-90% threshold. Therefore we use 0.5 as the default choice of prior. Difference grows larger as m increases.

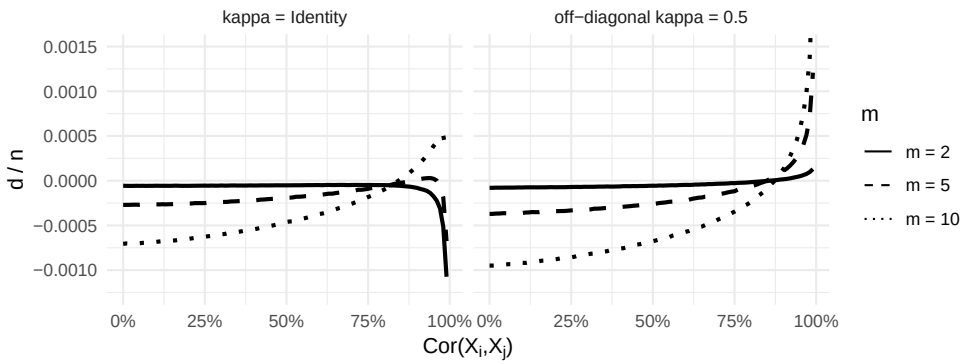


Figure 2: Difference d between SCC and DAG graph scores (divided by n data) as a function of true correlation between pairs of nodes (x axis) and number of nodes m. The hyperparameters are set to $q = m + 1$ and $\kappa = I_m$ in the left panel. In the right panel we change $\kappa_{ij} = 0.5$ for $i \neq j$.

The behaviour of d is also sensitive to variance of random variables. If κ is misspecified, the SCCs are always preferred for low variances and DAGs are always preferred for high variances. However, standardising the inputs can solve this problem, as will scaling κ by sampling variances of each node. This can be done automatically in software implementations and in both cases will "bring back" the behaviour of d to exactly what is seen in Figure 2.

As indicated by Equation 6, we can put a prior on correlation between two elements to any mean ρ by setting off-diagonal elements of κ to $(q - m - 1)\rho$, and any variance by adjusting q. In practice we can use this to manipulate the sign of d, thus allowing us to choose the level of correlation at which SCCs will be chosen over DAGs different from the 85%-90% threshold. This is illustrated in Figure 3.

3.2 MCMC algorithm for inference

The computer code needed to perform all of the examples has been implemented in the latest version of the *graph_sampler* software for MCMC inference on graphs, previously introduced by Bois and Gayraud (2015). Written in ANSI-standard C language, the full software is freely available at *www.nongnu.org/graphsampler* under the terms and conditions of the GNU General Public License, as published by the Free Software Foundation.

Graph_sampler uses Metropolis-Hastings algorithm to sample graphs from a scoring posterior distribution. The proposals in the algorithm are edge additions or

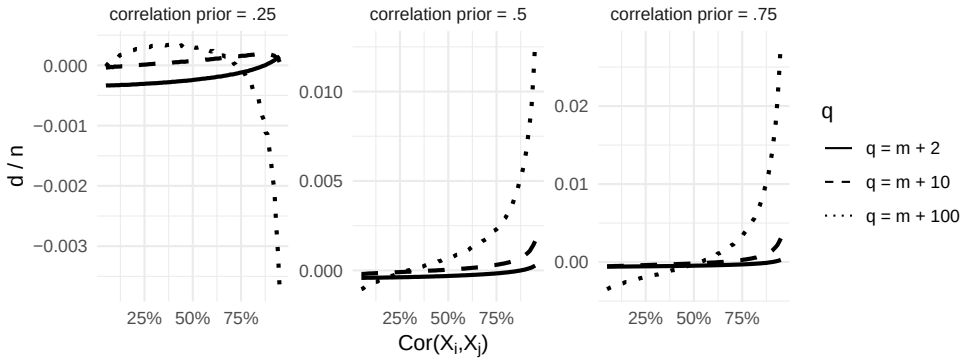


Figure 3: Difference d between SCC and DAG graph scores (divided by n data) under informative priors on correlation. κ is scaled according to q to yield desired mean correlation: 0.25 in the left panel, 0.5 in the middle, 0.75 in the right panel.

deletions, drawn according to Bernoulli prior on the graph edges. For DAGs, score of proposal is then evaluated by calculating difference in scores on the child node in the proposed addition or deletion. In all cases convergence to the target distribution can be checked by calculating Gelman-Rubin statistic (Gelman and Rubin, 1992) on chains of graph adjacency matrices. Convergence check function is included as part of the software.

When cyclical structures are allowed, the algorithm is modified to take into account situations where the condensed graph changes, (*i.e.* SCCs are created or deleted. Multiple nodes are affected in such situations and need to have their scores recalculated. We devised an additional decision rule to only condense graph (using Tarjan algorithm) when necessary and recalculate likelihood on the minimal set of nodes that may be affected by additions and deletions. It is presented in Supplement S3.

The MCMC approach yields a set of s graphs sampled from the posterior distribution. We represent them by their adjacency matrices $A^{(1)}, \ldots, A^{(s)}$. Such a sample can be used to approximate posterior probabilities of occurence of edges or motifs. For example, the probability of an edge from i to j, p_{ij}, is obtained by calculating $\hat{p_{ij}} = \sum A_{ij}/s$. However, such probabilities have to be treated with caution when cyclic structures are allowed. Depending on the objectives of analysis, we can either be interested in $\hat{p_{ij}}$ defined as above or the probability of i and j being part of the same SCC (p_{ij}^{SCC}) or of i being parent of j, but not in the same SCC (p_{ij}^{DAG}).

3.3 MCMC convergence in SCC setting

In MCMC algorithms for graph inference which operate only by adding or removing edges at each step, reversing the direction of an existing edge can be difficult. It requires two operations: a deletion followed by an addition. The first step will often (*e.g.*, in situations where two nodes are highly correlated) have an extremely low probability. Using tempered MCMC methods can solve this problem (see Barker

et al., 2010) but requires fine tuning of the tempering algorithm. Using SCCs provides a simpler solution: an addition (creating an SCC) followed by a deletion. Thus for some problems, allowing cyclic structures can be helpful even if we know that the true network is acyclic as it can avoid traversing these "probability wells".

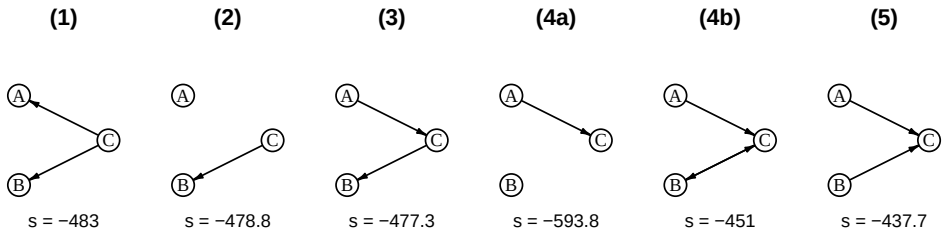


Figure 4: "Inverting a v-structure" in five steps. For each graph value s is $s(D|\mathcal{G})$. Leftmost graph is the starting graph for MCMC and the true generative mechanism is the rightmost. While deleting some edges such as CA can be "easy" in terms of log likelihood difference (steps 2 and 3), deletion of the CB edge (step 4a) is difficult as it leads into a "probability well". Adding the edge BC (4b), which results in an SCC, offers a way to reach (5) while avoiding the well.

We illustrate this with a simple example of inverting a v-structure. Assume $X_A \sim \mathcal{N}(0,1)$, $X_B \sim \mathcal{N}(0,1)$ and $X_C = X_A - 5X_B + \epsilon$ with $\epsilon \sim \mathcal{N}(0,1)$. We draw 100 realisations of each random variable. Assume the MCMC sampler starts from a graph model $A \leftarrow C \rightarrow B$ with score of -472.5. Assuming that we are working with DAGs only, any path to the true generating graph requires removal of CB edge. This is shown in Figure 4.

The well is a score difference of around 100 (therefore on average we will need e^{100} Metropolis-Hastings proposals to remove CB). Using SCCs easily circumvents this by creating an SCC involving the $B \leftrightarrow C$ SCC before deleting CB.

3.4 Linear model with additive noise

Our first example is straightforward, but designed to be difficult to correctly estimate. For this, we slightly expanded the graph from Figure 1 by adding node P, a parent to A, and Q, a child to E. We assumed a linear relationship and generated 100 draws for each node as follows: for j-th node, i-th generated value $x_j(i) = Pa_{x_j}(i) + \epsilon_{ij}$, where Pa_{x_i} are data for the parent node of X_j (for node P we set the mean to zero) and ϵ_{ij} are i.i.d. with $\mathcal{N}(0,5)$ for all i and j. For the SCC (nodes B, C and D) we used multivariate Gaussian distribution with same means and variances (equal to 5) and pairwise correlations of 0.9. This way, all of generated data was very highly correlated, making it difficult to distinguish between different graphs using likelihood alone, even under the correct assumption about data generating mechanism being a Gaussian linear additive noise model.

First, in the absence of prior information (Figure 5A) we did not succeed in retrieving the data generating graph and many superfluous edges were found. (Although we usually prefer to work with edge probabilities, for clarity of presentation we only show the best scoring graph here.) Including an informative prior on the out-degree

(power law with $\gamma = 3$) and size of SCCs (no larger than 3) enabled us to detect the SCC and the undirected edges correctly (Figure 5B). Lastly, adding information on the first cause, *i.e.*, enforcing (through a Bernoulli prior) $Pa(P) = \emptyset$, allowed us to retrieve the data generating graph (Figure 5C).

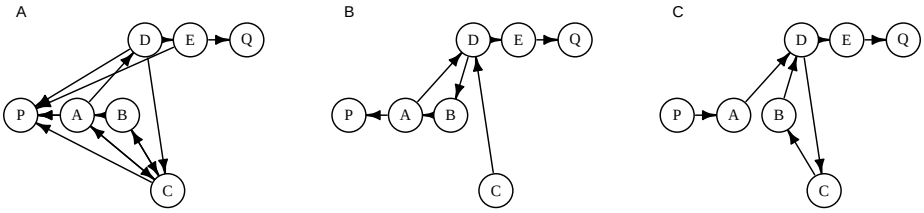


Figure 5: Graph structure inference when likelihood is the same as the generative model. Highest scoring graphs for: (A) A model without informative priors. (B) A model with prior on out-degree and SCC size. (C) A model with additional prior on the "first cause" (no edges allowed into node P).

The last two steps illustrate two difficulties with learning network structures. First comes the problem of detecting dependencies from data, which can be helped by putting a strong prior on the types of structures expected to occur in the graph (in this case degrees and SCC sizes). Even if we succeed in this, we are still left with multiple candidate graphs: in this case the condensed graph is a path from P to Q which is Markov-equivalent to a path from Q to P. Only the addition of a prior on whether P or Q is a probable cause can help us retrieve the true network.

As discussed, the choice between SCCs and DAGs is highly sensitive to correlation. We repeated the simulation using the same data-generating mechanism, but with a pairwise correlation of the SCC nodes equal to 0.5 instead of 0.9, DAGs were then preferred when using uninformative structural priors and the best graph resembled Figure 5(C) but without an SCC. Setting $q = m + 11$ and scaling the "default" κ appropriately is enough to bring up an SCC again.

3.5 SCCs detection in a 50-node linear model

In the second simulated study, we generated batches of 100 DAGs of 50 nodes by randomly permuting nodes and drawing each edge e_{ij} with probability of occurrence at 5% if $j > i$ (to avoid SCCs). We then created two three-node SCCs in each graph by adding all possible edges between two groups of three randomly selected nodes. Example of such graphs are presented in Figure 6. Each graph was then used as data generation mechanism for 100 data values for each node, according to a normal linear model with regression coefficients set to 1. That is, for j-th node, the i-th generated value $x_j(i) = \sum_{X_k \in Pa(X_j)} x_k(i) + \epsilon_{ij}$, with $\epsilon_{ij} \sim \mathcal{N}(0, 1)$ i.i.d. for all i and j. For SCCs the distribution was multivariate normal, with correlation between any two nodes fixed at 0.5 or 0.9, to benchmark performance in two different cases.

For inference, we compared four prior assumptions on the inverse Wishart parameters by varying q and setting κ to the desired correlation and scaling it appropriately

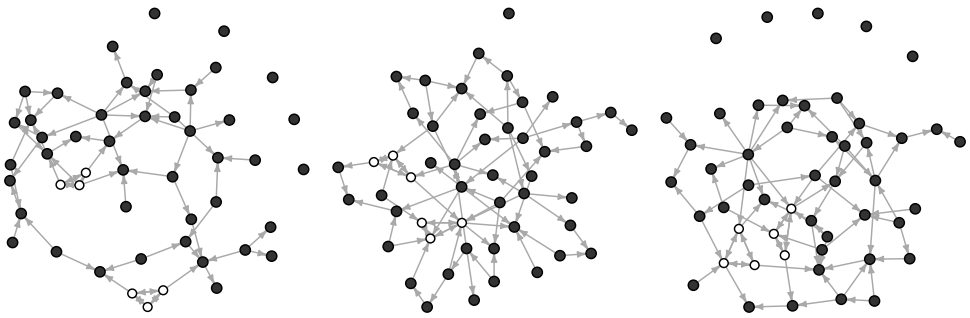


Figure 6: Examples of three graphs (out of 800) from which data was generated for this performance benchmark. All graphs were randomly generated with the same settings. White nodes belong to (three-node) SCCs.

(see Equation 6). We selected a "default" uninformative prior $q = m + 1$ assuming a within-SCC correlation of 0.5; a $q = m + 11$ prior (equivalent to 10 data points) assuming a correlation of 0.5; a $m+11$ prior assuming a correlation of 0.9; and a $m+101$ prior (equivalent to 100 data points) assuming a correlation of 0.5. A summary of these combinations is given together with results in Table 1.

In each case, we set a Bernoulli prior on the probability of edge occurrence $p(e_{ij}) = 0.025$ when $i \neq j$ (instead of 5%, as in graph generation half of off-diagonal edges were not allowed) and constrained the size of SCCs to be at most three, but imposed no more priors.

For each combination of "true" correlation and prior assumptions we generated data and ran MCMC inference 100 times. For each graph, we used 20 millions MCMC iterations, discarding first ten million, to infer on its structure. We only assessed MCMC convergence on a few selected graphs, but assumed that such run length was adequate given the simple nature of the problem and that we are interested in relative, not absolute, performance. For each of these runs, the probability of occurrence of SCCs was calculated from a sample of 100 adjacency matrices drawn from the MCMC chain. We report the area under the receiver operating characteristic curve (AUROC). It is the same as described in Marbach et al. (2010) – briefly, the k possible edges in the graph were ordered by probabilities obtained from MCMC and we calculated sensitivity and specificity k times, assuming that $1, 2, \ldots, k$ first edges occur and the rest do not. Note that perfect prediction (AUROC = 1) is impossible in this example, as we calculate our score for directed graphs and not equivalence classes. For SCCs, we only assessed sensitivity (as the AUROC statistic captures overall specificity well), by calculating a probability that the "true" SCCs are present in the MCMC results. Under our definition we needed to "detect" all three nodes of the SCC to count as a success. Table 1 presents results for both AUROC and "SCC sensitivity"; results are averaged over 100 inferences for each row.

Generally, the sensitivity and specificity (AUROC) of the score-based method is

Table 1: Sensitivity and specificity of the scoring method in retrieving true network structure.

True Correlation	Prior	AUROC	SCC Pr
0.50	q = m + 1; Cor = 0.5	0.93	0.12
0.50	q = m + 11; Cor = 0.5	0.91	0.26
0.50	q = m + 11; Cor = 0.9	0.86	0.01
0.50	q = m + 101; Cor = 0.5	0.91	0.52
0.90	q = m + 1; Cor = 0.5	0.89	0.10
0.90	q = m + 11; Cor = 0.5	0.89	0.40
0.90	q = m + 11; Cor = 0.9	0.87	0.95
0.90	q = m + 101; Cor = 0.5	0.91	0.39

good under this simple generative model. However, the detection of SCCs is low with "default" settings, with about 10% success rate. (Note that given that the equivalence class for three-node SCC is of size seven, *i.e.* SCC and six DAG configurations, so we would expect success rate of about 14% assuming equivalence of scores within class.) We can improve this by introducing informative priors. Generally highly correlated SCCs are easier to detect successfully, but using an informative prior helps. Misspecification of prior does not seem to overly impact the overall (AUROC) performance, but does affect the detection of SCCs. That is most salient in the case of a high correlation prior, as illustrated in the bottom left panel of Figure 7. When assuming a correlation of 0.50, the variability in success rate across graphs is large (with a peak around probability of 50%, corresponding to cases where one SCC has been identified perfectly and the other one not at all), but with a prior on correlation equal to 0.90 the behaviour is completely different.

4 DISCUSSION AND CONCLUSION

We proposed a model to represent cyclic structures within Bayesian networks. Our model offers an alternative way of describing joint probability distribution and performing network inference without apparent computational drawbacks. In our approach, SCCs are condensed to form multivariate nodes, which are still embedded in an acyclic Bayesian network. We can therefore factorise the likelihood, a key computational advantage of Bayesian networks. We use a score-based approach in a fully Bayesian setting. A posterior sample of graphs is obtained by MCMC sampling. This allows us to integrate prior knowledge on presence of edges, degrees, acyclic motifs, occurrence of SCCs *etc.* The placement of informative priors on network structure also brings faster convergence of MCMC sampling (if the data are not conflicting with the prior) by putting soft constraints of the size of the set of likely graphs.

The likelihood model we present is an additive linear model with Gaussian noise. Such model allows us to easily compute score by integrating out parameters. The only (arbitrary) constraint imposed by our Gaussian model is that the number of parents for all members of an SCC has to be less than the number of data points for each node. In the future other models for likelihood or other scoring functions should be explored. We also note that in the present form the model can only account for time

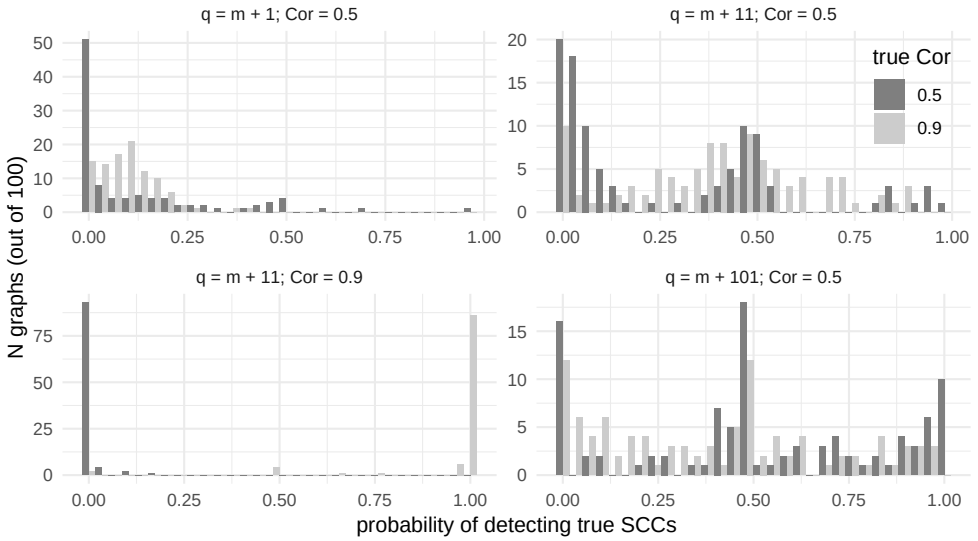


Figure 7: How the probability of detecting true SCCs changes with "true" correlation (differently coloured histogram bars) and four priors (panels of the figure). Peaks at 0.5 correspond to cases where only one of two SCCs has been detected.

as an additional linear term in regression, although a dynamic version of it might be workable. Use of SCCs with discrete random variables should also be explored.

Many alternative methods for characterising dependencies in graphs containing cycles have been proposed, including reciprocal graph models based on work by Koster (1996) (see paper by Ni et al., 2018, for recent application) or a heuristic algorithm approach to learning cycles from experimental data by Itani et al. (2010). Our work differs from statistical models for purpose of learning causal relationships in observational data, as under our model Markov-equivalent graphs can have different scores due to choice of priors and hyperpriors relating to SCCs.

Under the proposed model, detection of SCCs is sensitive to the choice of hyperparameters. Informative priors can be used to promote or suppress occurence of SCCs in the posterior. We can choose priors to favour SCCs over DAG structures even when correlation is lower than the threshold visible in Figure 2. This may be useful in applications where we know *a priori* that cycles are present or simply wish to describe joint distribution differently. However, in our model, a limitation in setting informative priors is tied to the properties of $\mathcal{W}^{-1}$ distribution, where the precision of variances and correlations is governed by a single parameter, q. Here again we must decide between standardisation and flexibility.

Our simulation study with a small linear model also shows the importance of prior choices on the inference. Note that pure likelihood-based inference amounts to placing only an indifferent Bernoulli prior on the adjacency matrix, and would bring the same inefficient inference as in Figure 5B. In the case of a larger network, the inference scales up well, but SCCs typically have a 50% chance to be detected. Finally, besides substantive applications, allowing for SCCs can reduce computation

time (by reducing the number of nodes) even for underlying DAGs, and improves convergence by easing edge reversals. The practical impact of those computational benefits should be explored in greater detail in the future.

Acknowledgements

F. Bois' work was funded by the Horizon 2020 project "EU-ToxRisk" of the European Commission (Contract 681002).

References

Acid, S., de Campos, L. M., Fernández-Luna, J. M., Rodríguez, S., María Rodríguez, J., and Luis Salcedo, J. (2004). A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service. *Artif Intell Med*, 30(3):215–232. 2

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8:450–461. 1

Alvarez, I., Niemi, J., and Simpson, M. (2014). Bayesian inference for a covariance matrix. *arXiv:1408.4050 [stat]*. arXiv: 1408.4050. 8

Barabási, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512. 33208. 4

Barker, D., Hill, S., and Mukherjee, S. (2010). MC(4): a tempering algorithm for large-sample network inference. In *Pattern Recognition in Bioinformatics*, volume 6282, pages 431–442. Springer-Verlag Berlin, Berlin. 10

Bois, F. Y. and Gayraud, G. (2015). Probabilistic generation of random networks taking into account information on motifs occurrence. *Journal of Computational Biology*, 22(1):25–36. 4, 9

Chickering, D. M. (2003). Optimal Structure Identification with Greedy Search. *J. Mach. Learn. Res.*, 3:507–554. 4

Cooper, G. F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Mach Learn*, 9(4):309–347. 2

Datta, S., Gayraud, G., Leclerc, E., and Bois, F. Y. (2017). Graph_sampler: a simple tool for fully Bayesian analyses of DAG-models. *Computational Statistics*, 32(2):691–716. 2, 4, 7

Geiger, D. and Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.*, 30(5):1412–1440. 7

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534. 8

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472. 10

Ghahramani, Z. (1998). Learning dynamic Bayesian networks. In *Adaptive Processing of Sequences and Data Structures*, Lecture Notes in Computer Science, pages 168–197. Springer, Berlin, Heidelberg. 2

Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Mach Learn*, 20(3):197–243. 04321. 7

Husmeier, D. (2004). Reverse engineering of genetic networks with Bayesian networks. *Biochemical Society transactions*, 31:1516–8. 1

Husmeier, D. and Werhli, A. (2007). Bayesian integration of biological prior knowledge into the reconstruction of gene regulatory networks with Bayesian networks. *Computational systems bioinformatics / Life Sciences Society. Computational Systems Bioinformatics Conference*, 6:85–95. 1

Itani, S., Ohannessian, M., Sachs, K., Nolan, G. P., and Dahleh, M. A. (2010). Structure Learning in Causal Cyclic Networks. In *Causality: Objectives and Assessment*, pages 165–176. 15

Koster, J. T. A. (1996). Markov properties of nonrecursive causal models. *Ann. Statist.*, 24(5):2148–2177. 15

Lauritzen, S. L. (1996). *Graphical Models*. Oxford Statistical Science Series. Oxford University Press, Oxford, New York. 05303. 1

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences of the United States of America*, 107(14):6286–6291. 13

Mukherjee, S. and Speed, T. P. (2008). Network inference using informative priors. *Proceedings of the National Academy of Sciences, USA*, 105(38):14313–14318. 2, 4

Ni, Y., Ji, Y., and Müller, P. (2018). Reciprocal Graphical Models for Integrative Gene Regulatory Network Analysis. *Bayesian Anal.*, 13(4):1095–1110. 00006. 15

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition. 1

Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM J. Comput.*, 1(2):146–160. 3

Verma, T. and Pearl, J. (1992). An Algorithm for Deciding if a Set of Observed Independencies Has a Causal Explanation. In *Proceedings of the Eighth International Conference on Uncertainty in Artificial Intelligence*, UAI'92, pages 323–330, San Francisco, CA. Morgan Kaufmann Publishers Inc. 4

Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J., and Jarvis, E. D. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics (Oxford, England)*, 20(18):3594–3603. 2

Zhou, X., Wang, X., Pal, R., Ivanov, I., Bittner, M., and Dougherty, E. R. (2004). A Bayesian connectivity-based approach to constructing probabilistic gene regulatory networks. *Bioinformatics (Oxford, England)*, 20(17):2918–2927. 2

Supplementary materials for "Structure learning of Bayesian networks involving cyclic structures"

We start by restating the main equation of the Section 2 in the main paper. The joint posterior distribution of both $(\theta, \Sigma)|(Y, X)$ can be expressed as

$$P(\theta, \Sigma|Y, X) \propto C_0 \cdot P(\theta|\Sigma, Y, X) \cdot P(\Sigma|Y, X) \tag{1}$$

where C_0, $P(\theta|\Sigma, Y, X)$ and $P(\Sigma|Y, X)$ are of the form:

$$\begin{aligned} C_0 =& (2\pi)^{\frac{m(k-n)}{2}} 2^{\frac{m(p-q)}{2}} \frac{\Gamma_m(\frac{p}{2})}{\Gamma_m(\frac{q}{2})} |\kappa|^{\frac{q}{2}} |X^t X|^{-\frac{m}{2}} |\kappa + A_0|^{-\frac{p}{2}}, \\ P(\theta|\Sigma, Y, X) =& \frac{|X^t X|^{\frac{m}{2}} |\Sigma|^{-k/2}}{(2\pi)^{\frac{mk}{2}}} \exp\left\{ \frac{-1}{2} tr\Sigma^{-1}(\theta - \hat{\theta})^t X^t X(\theta - \hat{\theta}) \right\}, \\ P(\Sigma|Y, X) =& \frac{|\Sigma|^{-(p+m+1)/2}}{2^{\frac{mp}{2}} \Gamma_m(\frac{p}{2})} \exp\left\{ -\frac{1}{2} tr\Sigma^{-1}(\kappa + A_0) \right\} |\kappa + A_0|^{\frac{p}{2}}, \end{aligned} \tag{2}$$

where $p = q + n - k$.

When θ and Σ are marginalized out from (1), we obtain(2), which corresponds to the prior predictive distribution of the strongly connected component under the *constant-Gamma* model, *i.e.*,

$$C_0 = s(\mathbf{x_j^L}|Pa(\mathbf{x_j^L})).$$

S1: EQUIVALENCE OF INVERSE GAMMA AND INVERSE WISHART SCORE

We will show that the p.p.d. obtained with the inverse Wishart distribution with $m = 1$ coincides with the inverse Gamma case. We use the same notation as in Section 2 of the paper. Recall that we model the distribution of $x_j|Pa(x_j)$ using the linear regression model with $m = 1$, X the design matrix and $Y = x_j$ and denote by θ and σ^2 the model parameters, where θ is a k-vector and σ^2 is a positive real number. Moreover, recall that the least square estimate $\hat{\theta}$ and the sample covariance matrix A_0 are given by: $\hat{\theta} = (X^t X)^{-1} X^t Y$ and $A_0 = (Y - X\hat{\theta})^t (Y - X\hat{\theta})$.

Then the likelihood is

$$L(\theta, \sigma^2; Y, X) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{1}{2\sigma^2}(Y - X\theta)^t (Y - X\theta)).$$

Let us consider the following independent priors on θ and σ^2:

$$\theta \propto Constant, \qquad \sigma^2 \sim InvGamma(q/2, \tau/2).$$

The p.p.d. is then obtained from

$$s(Y|Pa(Y)) = \int \frac{\tau^{q/2}}{2^{q/2}\Gamma(q/2)(\sigma^2)^{q/2+1}} \exp(-\frac{\tau}{2\sigma^2}) \left(\int L(\theta, \sigma^2; Y, X) d\theta \right) d\sigma^2.$$

First, integrating out θ leads to

$$\int L(\theta, \sigma^2; Y, X) d\theta = \frac{(2\pi\sigma^2)^{k/2} \ |X^t X|^{-1/2}}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\hat{\theta})^t (Y - X\hat{\theta}) \right)$$

Second, integrating out σ^2 leads to $s(Y|Pa(Y))$; indeed,

$$\begin{aligned} s(Y|Pa(Y)) &= \frac{\tau^{q/2}(2\pi)^{k/2}|X^t X|^{-1/2}}{2^{q/2}(2\pi)^{n/2}\Gamma(q/2)} \times \frac{2^{\frac{n+q-k}{2}}\Gamma(\frac{n+q-k}{2})}{(\tau + A_0)^{\frac{n+q-k}{2}}} \\ &\quad \times \int \frac{(\tau + A_0)^{\frac{n+q-k}{2}}}{2^{\frac{n+q-k}{2}}\Gamma(\frac{n+q-k}{2})} \frac{1}{(\sigma^2)^{\frac{n+q-k}{2}+1}} \\ &\quad \exp\left(-\frac{1}{2\sigma^2}(\tau + A_0) \right) d\sigma^2 \\ &= (2\pi)^{\frac{k-n}{2}} \ 2^{\frac{n-k}{2}} \ \frac{\Gamma(\frac{n+q-k}{2})}{\Gamma(q/2)} \tau^{q/2} \ |X^t X|^{-1/2} \\ &\quad (\tau + A_0)^{-\frac{n+q-k}{2}}, \end{aligned} \tag{3}$$

which corresponds to the p.p.d. for the SCC case defined in Section 2 of the paper, assuming $m = 1$.

S2: PRIOR PREDICTIVE DISTRIBUTION FOR TWO-NODE GRAPHS

We will now show that the p.p.d.'s for two-node graphs are not equal. We denote the nodes by A and B and focus on p.p.d.'s for an empty graph $\mathcal{G}_1$, SCC graph $\mathcal{G}_2$ ($A \leftrightarrow B$) and DAG $B \to A$, $\mathcal{G}_3$. All are evaluated under the same data $D = (x_A, x_B)$.

We denote by 1_n a column vector of ones of dim n, $\bar{x}_A$ and $\bar{x}_B$ are the empirical mean of x_A and x_B respectively while s_A^2 and s_B^2 are the empirical variances of x_A and x_B; $s_{A,B}$ denotes the empirical covariance between x_A and x_B and $\sigma_{A,B}$ denotes the empirical variance covariance matrix of D.

Empty graph: The p.p.d. for the empty graph is equal to $s(x_A|Pa(x_A))\times s(x_B|Pa(x_B))$, where $Pa(x_A) = Pa(x_B) = \emptyset$ and each term is given by Equation (3) with $X = 1_n$, $|X^tX| = n$, $A_{0,B} = (x_B - 1_n\bar{x}_B)^t(x_B - 1_n\bar{x}_B)$ and $A_{0,A} = (x_A - 1_n\bar{x}_A)^t(x_A - 1_n\bar{x}_A)$; we then have,

$$s(x_A, x_B|\mathcal{G}_1) = (2\pi)^{1-n}\, 2^{n-1} \left(\frac{\Gamma(\frac{n+q-1}{2})}{\Gamma(\frac{q}{2})}\right)^2 \frac{\tau^q}{n} [(\tau + ns_B^2)(\tau + ns_A^2)]^{-\frac{n+q-1}{2}}.$$

SCC graph: The p.p.d. for $\mathcal{G}_2$ is given by C_0 defined in (2) with $m = 2$, $k = 1$ since $Pa(x_A, x_b) = \emptyset$ and $A_0 = \begin{pmatrix} x_A - \bar{x}_A 1_n & x_B - \bar{x}_B 1_n \end{pmatrix}^t \begin{pmatrix} x_A - \bar{x}_A 1_n & x_B - \bar{x}_B 1_n \end{pmatrix}$, that is $A_0 = n\sigma_{A,B}$. Finally, the score is

$$s(x_A, x_B|\mathcal{G}_2) = (2\pi)^{-\frac{2(n-1)}{2}} 2^{\frac{2(n-1)}{2}} \frac{\Gamma_2(\frac{n+q-1}{2})}{\Gamma_2(\frac{q}{2})} |\kappa|^{\frac{q}{2}} |n|^{-\frac{2}{2}} |\kappa + A_0|^{-\frac{n+q-1}{2}}.$$

DAG graph: The p.p.d. for $\mathcal{G}_3$ is defined by $s(D|\mathcal{G}_3) = s(x_A|x_B) \times s(x_B|Pa(x_B))$ where $Pa(x_B) = \emptyset$.

Score $s(x_B|Pa(x_B))$ is given by Equation (3) with $X = 1_n$, $|X^tX| = n$, and $A_{0,B} = (x_B - 1_n\bar{x}_B)^t(x_B - 1_n\bar{x}_B)$.

The score $s(x_A|x_B)$ is defined by Equation (3) with

$$\begin{aligned} k &= 2, \\ X &= \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{B,1} & x_{B,2} & \dots & x_{B,n} \end{pmatrix}^t, \\ X^tX &= \begin{pmatrix} n & n\bar{x}_B \\ n\bar{x}_B & \sum_i x_{B,i}^2 \end{pmatrix}, \\ X\hat{\theta} &= \left(\left((\bar{x}_A + (x_{B,i} - \bar{x}_B)\frac{s_{A,B}}{s_B^2}) \right)_i \right), \\ A_0 &= ns_A^2 - n(s_{A,B}^2/s_B^2). \end{aligned}$$

Combining the above terms provides the p.p.d. for $\mathcal{G}_3$:

$$s(x_A, x_B|\mathcal{G}_3) = (2\pi)^{-\frac{2n-3}{2}} 2^{\frac{2n-3}{2}} \frac{\Gamma(\frac{n+q-1}{2})\Gamma(\frac{n+q-2}{2})}{(\Gamma(\frac{q}{2}))^2} \tau^q \frac{|\tau + ns_B^2|^{-\frac{n+q-1}{2}}}{n^{1/2}} \frac{|\tau + ns_A^2 - n\frac{s_{A,B}^2}{s_B^2}|^{-\frac{n+q-2}{2}}}{(ns_B^2)^{1/2}}.$$

This expression depends on s_B^2 and s_A^2 in a way that doesn't allow for equivalence between $A \to B$ and $B \to A$.

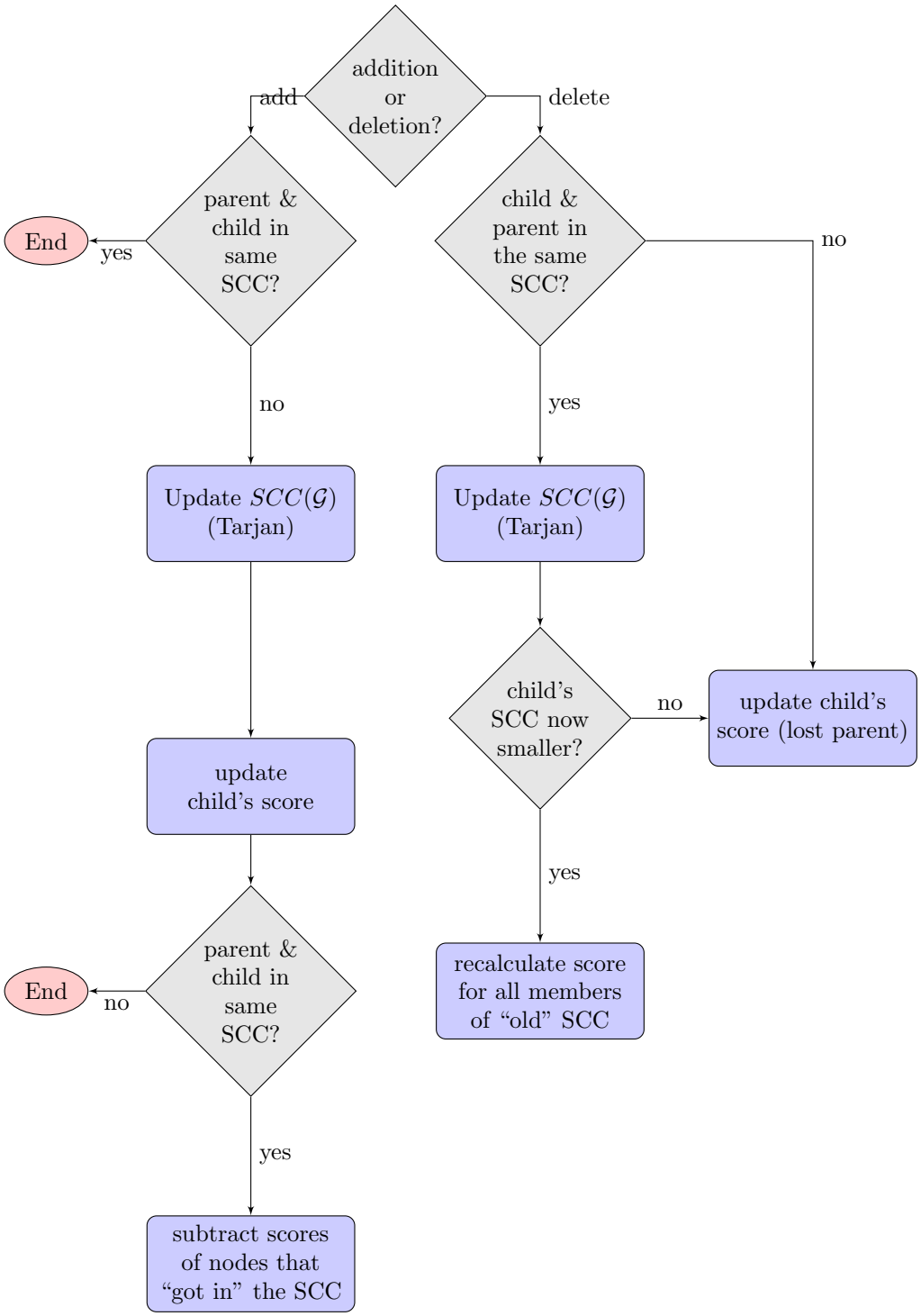
S3. ALGORITHM FOR UPDATING SCORE IN SCC CASES

We present here a simplified flowchart for updating the graph score (marginal likelihood) when edges are added or deleted with minimum of necessary re-calculations. By "updating" score we mean calculating a difference between steps, which then allows us to evaluate if the jumping proposal is accepted (and if yes, to update the score function).

In a DAG case only the child's conditional probability is affected when we change edges. Therefore from step to step we only need to store a vector of individual node scores, a proposed score for child, and the current score (sum of all scores)[1]. If proposal is accepted, we update the score by a difference in child scores before and after.

In a cyclic graph the incremental update of score is more difficult. We need to store additional structure describing SCCs: their sizes and member nodes for each. An addition of an edge may result in "closing a loop" and deletion may "destroy" an SCC. Moreover, in case of addition we may be creating one larger SCCs out of two smaller SCCs while in the case of removal a new, smaller SCCs may appear where a bigger one was deleted. In the worst scenario score of all nodes can change via a single addition or removal. We avoid this where possible by using a set of simple *if-else* rules outlined below.

[1] In practice we also store and dynamically update parent sets and their sizes for each node, as this information is used every time score is evaluated.


addition or deletion?
add
delete
parent & child in same SCC?
yes
End
no
Update $SCC(\mathcal{G})$ (Tarjan)
update child's score
parent & child in same SCC?
no
End
yes
subtract scores of nodes that "got in" the SCC
child & parent in the same SCC?
no
yes
Update $SCC(\mathcal{G})$ (Tarjan)
child's SCC now smaller?
no
update child's score (lost parent)
yes
recalculate score for all members of "old" SCC

2.2 Implementation of the structure learning MCMC algorithm in C and R

As described in the Section 3.2 of the methodological paper, the MCMC algorithm for inference on Bayesian networks containing cyclical structures is implemented in the *graph_sampler* software. This modification has been released as a part of version 3.0 of *graph_sampler*. The user manual and the complete program code (distributed under a GNU GPL license) can be found at `http://www.nongnu.org/graphsampler/`.

Subsequently, we developed a software package that implements *graph_sampler* directly within R. The package is called *rgraphsampler* and is available on-line at `https://github.com/wwiecek/rgraphsampler`. The development of this additional piece of software was motivated in two ways: one, to make available to R users a new and independent tool for performing Bayesian network inference and two, to make working with outputs of *graph_sampler* analyses easier. Full documentation for *rgraphsampler* is included on-line and here we only recapitulate the basic functionality included in the package.

R version is fully featured in that it directly uses unmodified *graph_sampler* (version 3.0) code within the package. However, GSL versions of functions are disabled to decrease dependencies for non-Linux users. Therefore the results obtained through R version should be equivalent to running non-GSL version of the software.

The main function of the package is `rgs()`, which calls *graph_sampler* and, if requested, automatically loads outputs into R. Input scripts for *graph_sampler* cannot be generated directly by *rgraphsampler*; however, 30 scripts are included in the package for testing and demonstration. Similarly, `rgs()` always writes to disk the standard output files, which are exactly the same as when running *graph_sampler*. However, the outputs (adjacency matrix samples, highest scoring graph, probability of edges, list of strongly connected components) are automatically parsed and loaded into R, unless the user requests otherwise.

The R package also includes rudimentary functions for simulating data according to a given causal network (under the assumption of linear Gaussian model); cyclical structures are also allowed. This functionality is useful for model criticism, as the fit to simulated data can be assessed quickly when building models. Basic plots (based on `igraph` package) and text summaries are also included and produced automatically in R. A short example for this functionality is given in the package vignette, which recreates the simple case study of the 7-node graph which is presented in the methods paper.

Chapter 3

Meta-analysis model for risk assessment of chemicals

The work presented in this chapter is based on two articles. In Wiecek et al. (2019b) we described the model and provided some case studies to illustrate these modelling considerations. Previously, in Quignot et al. (2018), we applied a version of the model to characterise metabolism of mixtures of chemical compounds in a 3A4 enzyme.[1] Both publications are presented here, together with Stan codes for models described in the methods paper.

[1] For the latter, the author contributed the statistical model, conducted statistical inference and co-wrote the methods part of the paper.

A generic Bayesian hierarchical model for the meta-analysis of human population variability in kinetics and its applications in chemical risk assessment

Witold Wiecek, Jean-Lou Dorne, Nadia Quignot, Camille Bechaux, Billy Amzal

July 2019

Abstract

Traditionally, chemical risk assessment in humans aims to derive safe levels of exposure to chemicals using toxicological data in test species while applying a default 100-fold uncertainty factor (UF) to allow for both interspecies differences (10) and human variability (10) in toxicokinetic (TK) and toxicodynamic (TD) processes. Over the last two decades, meta-analysis methods have allowed to perform quantitative analysis of population variability in kinetics and dynamics to replace these default UFs with data-based UFs. Pathway-related UFs can now be derived using population variability in kinetics for known metabolic pathways. This study presents a new hierarchical Bayesian model for the meta-analysis of human population variability in TK parameters, along with its application to the derivation of UFs. Data standardisation and data gaps are also discussed, as well as model implementation, selection and validation. Applications of the generic model in human risk assessment are illustrated through a case study quantifying inter-phenotypic differences in TK for substances metabolised via the CYP2D6 polymorphic enzyme. Model implementations in open-source software are provided and future refinements and applications of the model are proposed.

Keywords: meta-analysis, Bayesian hierarchical model, human risk assessment, uncertainty factors, variability

Highlights

- A hierarchical Bayesian model for the meta-analysis of human kinetic data has been developed in R.
- Using the model, uncertainty factors for chemical risk assessment can be derived.
- Applications are described for modelling in generic subgroups of human populations and polymorphic CYP2D6 metabolism.
- Broader applications in chemical risk assessment are discussed.

1. Introduction

Human health risk assessment has been shifting from qualitative to quantitative approaches as part of a "weight of evidence" (WoE) approach, defined by the WHO as a "process in which all of the evidence considered relevant for a risk assessment is evaluated and weighted" [1]. WoE approaches select, weigh and integrate the evidence in a systematic, consistent and transparent way [2,3]. In the context of chemical risk assessment, hazard characterisation sets safe levels of chemical residues in food to protect human health (e.g. tolerable daily intake (TDI) for contaminants). These safe levels have been traditionally derived by applying a default uncertainty factor (UF) of 100 to a reference point or point of departure from chronic to sub-chronic studies in test species (rat, mice, dog, rabbit) [4]. The default factor of 100-fold allows for interspecies differences (10-fold) and human variability (10-fold) [5]. In the 1990s, the UFs for interspecies and human variability have been subdivided to account for the differences in toxicokinetics (TK) and toxicodynamics (TD). For humans, two equal default uncertainty factors ($\sqrt{10} = 3.16$) have been proposed. These default values can be replaced by data-derived UFs or chemical-specific adjustment factors when chemical-specific data are available. For the kinetic dimension, Renwick and Lazarus proposed to use pathway-related default UFs [6].

Meta-analyses on human variability in TK for markers of acute (Cmax) and chronic exposure (AUC and clearance), based on the therapeutic drug databases, have proposed pathway-related default UFs [7–14]. These studies have highlighted large variability in individuals for polymorphic phase I (CYP2C9, CYP2C19 and CYP2D6) and phase II (NAT2) metabolism [15]. However, data for individual phenotypes were limited at the time in terms of number of studies and sample size (typically below 10). Over the last decade, significant new data have been published on the impact of phenotypes on the kinetics of therapeutic compounds. Additionally, the previously published meta-analyses were based on weighted averages assuming fixed effect models with inverse variance weights [8] and did not address the relative contribution of the variability across subgroups to the overall variability in the datasets, leading to uncertainty in the parameter estimates.

Recently, meta-analysis methods for health-care and chemical risk assessment have been further developed using Bayesian approaches to allow for quantification of variability and uncertainty, particularly for datasets with large heterogeneity across studies and models with complex hierarchical structure [16,17]. Bayesian models can characterise uncertainty and variability better than simple meta-analysis methods by partitioning observed variance between sampling variation, heterogeneity across studies, and other sources of variability, e.g. between subgroups of population. This, in turn, gives more precise estimates across compounds and provides means to extend the inference to unobserved compounds and subgroup combinations. Consequently, a generic model which is able to "pool" these estimates across studies, substances and subgroups of human populations, while also accounting for sample variation, is much needed in this area.

Herein, a new hierarchical Bayesian model for the meta-analysis of kinetic data in subgroups of human populations is presented. Applications of the generic model are one, quantifying inter-individual differences in kinetics between subgroups of human populations; two, quantifying inter-phenotypic differences in kinetics for polymorphic enzyme metabolism. From such model UFs for chemical risk assessment can be derived. Data collection and model validation methods are also discussed, followed by an illustrative case study for polymorphic CYP2D6 metabolism. Future refinements and applications of the hierarchical Bayesian model with regards to the meta-analysis of kinetic data for specific metabolic pathways, the calibration of quantitative *in vitro in vivo* extrapolation models (QIVIVE) in humans, as well as the meta-analysis of kinetic and toxicity data in species of ecological relevance are considered.

2. Methods

This section presents the Bayesian meta-analysis model. Data collection and associated methodological issues are discussed first. The description of the statistical model for the meta-analysis of human kinetic data follows, including how to use individual-level data. The generic model allows for comparison between a reference group (e.g. healthy adults) and subgroups of human populations (e.g children, neonates, elderly). A modification of the generic model is then developed for the analysis of human inter-phenotypic differences in kinetics for polymorphic metabolism. Derivation of uncertainty factors, model validation and model selection are also discussed.

2.1. Data collection of human kinetic parameters in subgroups of the population

Meta-analysis of kinetic data requires sample means and sample variations reported by individual studies, classified by compound and kinetic parameter. Such data are obtained through systematic reviews or extensive literature searches. From a data collection perspective, aggregate data are reported using a range of measures, regardless of whether a geometric or an arithmetic scale is used. For dispersion, commonly reported statistics are standard deviation, standard error, 95% confidence interval, inter-quantile range or min-max range. Therefore, before conducting a meta-analysis data must be converted to a common format.

Kinetic data are assumed to be log-normally distributed. Log-normal distributions are appropriate in multiplicative models which do not show any negative values and have been applied to many biological processes including body weights, particle sizes, tolerance to drugs, systolic and diastolic blood pressure and pharmaco/toxicokinetic data; in contrast, normal distributions are appropriate in additive models [6,8,18–21]. Therefore, all measures of dispersion are converted to geometric standard deviation (gsd) and all means or medians to geometric means (gm). For the purpose of modelling, all values are further transformed into logarithms to obtain normal distributions, i.e. $lgm = \log(gm)$ and $lv = \log(gsd)^2$.

In some publications, ratios of geometric means are reported alongside subgroup-specific means. However, since ratios can be derived from means, it is not necessary to collect them. While they are not used as inputs for the Bayesian meta-analysis model, they are important to model validation and will be addressed later in this section.

While conducting systematic reviews and meta-analysis of kinetic data, common key considerations include:

Sample size considerations

Sample sizes in human kinetic studies are often relatively low, most often between 3 and 10. Therefore, for the modelling of such datasets, distribution of sampling variance should be taken into account. In addition, where variance is estimated from range data and medians, adjustment for small sample sizes can be made using estimators proposed by Hozo *et al* [22].

Availability of individual-level data

While kinetic studies with small sample sizes typically do not report individual-level data, they do report "summary" statistics; e.g. studies with sample size of 3 often report median, minimum and maximum, which are *de facto* all individual-level values. A method to include individual-level data is presented later in this section.

Errors in reporting dispersion

In addition to reporting many different dispersion statistics, some studies of kinetic data report the standard deviation (SD) as the standard error (SE) or provide the 95% intervals calculated using SD rather than the SE. Such erroneous reporting may affect the overall estimates of variability, even when it occurs in only a few studies in the dataset. Each study should be checked carefully before running the meta-analysis.

Correlated data

A high proportion of kinetic studies report multiple values of the same parameter for a given subgroup of the human population (polymorphism). This may be due to to repeated interventions in the same study subjects or interventions in multiple groups. When the proportion of multiple values is high, additional step should be taken to prevent treating correlated values as independent. One solution is to meta-analyse values on a study-subgroup level, resulting in a single pair of logarithm geometric means and variance (lgm and lv) per each study-subgroup pair. This can be done either through a simple modification of the model presented in the next section or prior to analysis. In some cases, when multiple values are reported, additional information may be available that will allow to stratify reported means and variances by experimental condition or characteristics of the cohort; such cases could be considered as separate experiments.

2.2. Hierarchical Bayesian model for the meta-analysis of human kinetic data

The Bayesian meta-analysis model presented here aims to quantify how population means and variances vary across subgroups of human populations for kinetic data. The model is concerned with behaviour of means and variances on parameter-, subgroup- and compound-specific basis. For simplicity, all collected data are assumed to be for a single parameter. Indices $i = 1, 2, \ldots, N$ are for the reported logarithms of sample geometric means, lgm, the reported logarithms of sample variations, lv, and the respective sample sizes, denoted as n. As mentioned above, the inputs for the meta-analysis are the sample means and variations lgm and lv. These depend on true study means μ_i and true study standard deviations σ_i through sampling distributions:

$$lgm_i \sim \mathcal{N}(\mu_i, \frac{\sigma_i^2}{n_i})$$

$$lv_i \sim \Gamma(\frac{n_i - 1}{2}, \frac{n_i - 1}{2\sigma_i^2})$$

The Gamma distribution above is parameterised as (shape, rate).

Inter-individual differences between subgroups of human populations

The generic model allows to quantify inter-individual differences in kinetics as ratios of geometric means R (relative to a reference group). For observation i, the index of study, compound and subgroup of the population associated with that observation are $s(i)$, $c(i)$ and $g(i)$ respectively; such indices are written in subscript. A particular group g_1 is chosen as reference. This is done on a case-by-case basis, depending on the context of the meta-analysis (e.g. healthy adults). True means (μ_i) and true variances (σ_i^2) are assumed to depend on parameters specific to a given study, compound and subgroup. For μ_i, it is assumed that there is a separate (fixed) baseline value for each compound (μ^c), randomly affected by study effect (μ^s). That is, $\mu^s \sim \mathcal{N}(0, (\sigma^s)^2)$, where σ^s measures the variability of means across studies (common for all substrates). Ratios of geometric means between different subgroups are denoted by R. For σ_i, a dependence on the compound (through parameter γ^c) and the subgroup (parameter γ) on a log scale is assumed:

$$\begin{aligned} \mu_i &= \mu^c_{c(i)} + \mu^s_{s(i)} + log(R_i), \\ log(\sigma_i) &= \gamma^c_{c(i)} + \gamma_{g(i)}. \end{aligned}$$

Since most kinetic studies in the literature have small sample size and reflect population variability, the observed ratios between subgroups are highly variable. Therefore, the ratios are treated as random variables:

$$log(R_i) \sim \mathcal{N}(\mu^g_{g(i)}, (\sigma^g)^2),$$

where $\mu^g_{g(i)}$ is the mean log-ratio for group $g(i)$ and σ^g is its variability across studies. For the reference group, $\mu^g_{g_1} = 0$ and $\gamma_{g_1} = 0$. While the default assumption is that R varies across studies, the model can be modified to assume fixed ratio R by setting $\sigma^g = 0$.

For each compound, priors for both the mean value μ^c and the mean variability γ^c are independent, with $\mu^c \sim \mathcal{N}(0, 10^2)$, $\gamma^c \sim \mathcal{N}(0, 5^2)$. The model uses a half-normal prior for σ^s and σ^g, with $\sigma^s \sim \mathcal{N}^+(0, 5^2)$ and $\sigma^g \sim \mathcal{N}^+(0, 2.5^2)$. When the number of groups is low (e.g. 3) or convergence is poor, other priors for σ should be considered; appropriate alternatives are discussed in the literature [23]. For γ, a $\gamma \sim \mathcal{N}(0, 5^2)$ prior is used.

The quantities in this section, apart from lgm, lv and sample size, are all model parameters. Statistical identifiability may be an issue where data are sparse. A simpler model, may be preferable, particularly for variations, where γ^c can be replaced by a single baseline value shared by all compounds. Similarly, in some settings either no studies are available with two subgroups directly compared or the variability of the ratios is low, justifying fixing $\sigma^g = 0$ as mentioned above. In such cases, a choice of whether to include different parameters can be made through the default model selection approach proposed below

Analysing inter-phenotypic differences in kinetics for polymorphic enzyme metabolism

The generic model can be modified to allow the quantification of the inter-phenotypic differences in polymorphic enzyme metabolism on human kinetic parameters. In what follows, three groups within human populations are taken into consideration: extensive metabolisers (EM), as the reference group, intermediate metabolisers (IM) and poor metabolisers (PM). The model can be extended to more groups when modelling different enzymes.

The impact of enzyme polymorphism on kinetic parameters is determined by two factors: the degree of the enzyme's functional impairment (EF) and the fraction of the compound (on a percentage of the dose basis) metabolised by the enzyme, fm. EF of 1 means perfect function and 0 is complete impairment. EM is assumed to be complete function and used as the reference group, while PM is assumed to be complete impairment. For IM, the mean EF is an unknown value between 0 and 1. The ratios in PM relative to the reference group EM for kinetic parameters such as clearance (CL) or area under the plasma concentration curve (AUC) are given by

Eq (1)

$$R_{CL} = \frac{CL_{PM}}{CL_{EM}} = EF \cdot fm + (1 - fm),$$
$$R_{AUC} = 1/R_{CL}.$$

Owing to this definition of R, fm can be calculated directly from measurements of CL or AUC values, under the assumption that the mean enzyme function is equal to 0 for PM subgroups and equal to 1 for all EM subgroups. In that case $fm = 1 - R_{CL}$. However, there is a large variation in genotypes within the phenotype. In CYP2D6, PM phenotypes are associated with homozygosity for alleles with no activity (i.e.*3,*4,*5,*6,*7,*8,*11,*14,*15,*18,*19,*20,*21,*40); IM phenotypes with either homozygosity with decreased activity alleles (i.e.*9,*10,*17,*21,*36,*29,*41,*45,*46) or heterozygosity with one decreased activity alleles and one non activity alleles; and EM phenotypes with either homozygosity with normal activity alleles (i.e.*1,*2) or heterozygosity with one decreased activity alleles and one non activity alleles.

In the case of polymorphic subgroups, the modification of the generic model is as follows: R is assumed to be dependent on compound c and polymorphism p (through EF), in contrast to the generic model, where R was not compound-specific. EF itself is assumed to vary across studies. In the case of CL, the ratio of PM to EM is given as

$$R_i = EF_i \cdot fm_{c(i)} + 1 - fm_{c(i)}.$$

To make the model identifiable, the two extreme means are fixed at 0 and 1, i.e. $EF_{EM} = 1$ and $EF_{PM} = 0$, while EF_{IM} is treated as a parameter to estimate. Another parameter to estimate is variability in EF's across studies.

Assuming that some additional *in vitro* data on fm are available for some compounds in addition to *in vivo* kinetic data, they can be used to construct priors on these particular compounds, as will be demonstrated in the case study. Where no *in vitro* data are available, a vague prior or expert knowledge can be used to construct fm priors. For all variables in $[0, 1]$ range, truncated normal distributions are used, denoted as $\mathcal{N}_{[0,1]}$. Indexing over observations by i,

$$\begin{aligned}
EF_{EM}(i) &\sim \mathcal{N}_{[0,1]}(1, (\sigma^{EF})^2) \\
EF_{IM}(i) &\sim \mathcal{N}_{[0,1]}(\mu_{IM}^{EF}, (\sigma^{EF})^2) \\
EF_{PM}(i) &\sim \mathcal{N}_{[0,1]}(0, (\sigma^{EF})^2) \\
\sigma^{EF} &\sim \mathcal{U}(0, 1) \\
\mu_{IM}^{EF} &\sim Beta(2, 2)
\end{aligned}$$

As discussed in the case of the generic model, in particular cases it may be useful to treat ratios as fixed (without introducing the variability between studies). This can also be achieved in this model, by fixing $\sigma^{EF} = 0$ and using vague priors on fm.

A schematic presentation of the Bayesian hierarchical model, both for the generic case of subgroups of human populations and for polymorphic metabolism, is presented in Figure 1.

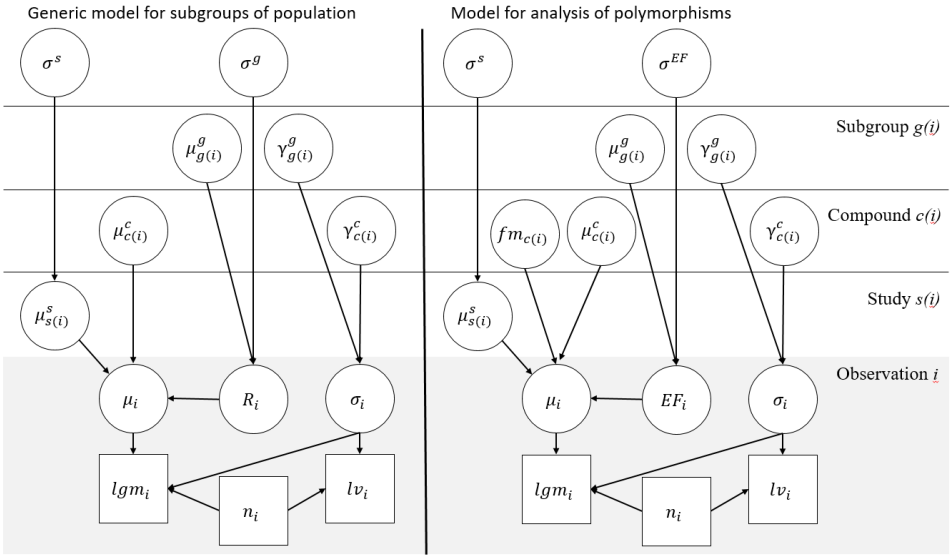


Figure 1. Scheme of the Bayesian model for the meta-analysis of kinetic data: in subgroups of human populations (left panel) and for polymorphic metabolism (right). Squares represent data, circles – random variables and arrows denote dependencies. With exception of the two parameters at the top (variability of means across groups and studies), all variables are multi-dimensional: observations are indexed by i and depend on study-, compound- and subgroup-specific variables, each with their specific set of indices.

Use of individual-level data in meta-analysis model

The generic model can also be applied to situations when individual level observations are available. If for a given study-arm i, the individual-level observations $y_i = (y_{i,1}, \ldots, y_{i,n_i})$ are available for j subjects, the aggregate-level contribution of lgm_i and lv_i to the overall likelihood can be replaced with the likelihoods of these individual values. For the j-th subject,

$$\log(y_{i,j}) \sim \mathcal{N}(\mu_i, \sigma_i^2),$$

owing to the assumption of log-normality of kinetic parameters. This approach is easy to implement in Markov Chain Monte Carlo (MCMC) software as no new parameters or "re-indexing" of other vectors are introduced. The simplicity of implementation owes to the base model treating both μ_i and σ_i as unknown parameters. The model presented in the previous section remains valid even in the presence of individual-level data only.

2.3. Derivation of uncertainty factors for chemical risk assessment

The results of the meta-analysis of kinetic data can be used to derive UFs for chemical risk assessment. $UF(x)$ is defined as the ratio of the x-th percentile of the population to the median in the reference group. Often the 95th percentile is used [24]. When the ratio between the subgroup and the reference group is larger than the ratios of quantiles within the subgroups, $UF(x)$ is given as the ratio of x-th percentile in the group of interest to the median in the reference group. In that situation, the UF can be derived based on two inputs only: the ratio R and the associated variation in the subgroup.

For situations where the variation within the reference group is high or a risk assessment is needed for a particular population, a simulation approach can be taken. Distribution of different subgroups in the population of interest (denoted p) is required, as well as N, the size of the population. (For theoretical cases N can be set to a large number, e.g. 10,000.) $UF(x)$ is then derived using the following algorithm:

1. Draw N values according to distribution into groups, p
2. For each N, draw a value according to the model's mean and variance distributions for that particular group, yielding an N-dimensional vector S
3. Calculate $\hat{UF}(x)$ as a ratio of x-th percentile of S to the median of S

This procedure is repeated M times, yielding M estimates of $\hat{UF}$, with M set to at least 10,000. Variation in $\hat{UF}$ reflects the model uncertainty and can be reported together with the mean $\hat{UF}$.

2.4. Model selection and validation

The statistical fit of the model to data should be assessed visually on both means and variances, using a plot comparing observed and expected values. Since both μ and σ are Bayesian parameters, they can be plotted with uncertainty intervals (e.g. 95%). For the means, a lack of fit to the observed values may occur when the observed ratios are highly variable. For the variances, a weaker fit can be expected under the proposed model, since no random effect is used by default. In both cases a visual check for systematic error in particular subgroups (especially for μ) and compounds (especially for σ) is required. Additionally, the fit of estimated to observed ratios of geometric means should be assessed separately, since these are the most relevant quantities for the derivation of UFs. The ratios are derived on a study-by-study basis from the geometric means in the data table of inputs.

Model selection can be implemented through cross-validation, by measuring the ability of a given model to correctly predict values of observations which were not used. A generic metric of predictive performance is log predictive density (lpd). In a model

with fixed R, lpd for J new observations of lgm and lv, indexed by $j = 1, \ldots, J$, lpd is given as

$$lpd = \sum_{j \in J} \big(\log(\Gamma_{\text{pdf}}(lv_j | \frac{n_j - 1}{2}, \frac{n_j - 1}{2\sigma_j^2})) + \log(\Phi(lgm_j | \mu_j, \frac{\sigma_j}{\sqrt{n_j}})) \big)$$

where $\Gamma_{\text{pdf}}(x|k, \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x\theta}$ and Φ is probability density function of the normal distribution; μ_j and σ_j are defined as above. Since these are random variables, lpd is also a random variable and approximated by drawing values of all parameters (using their conditional distributions).

As a default, it is recommended to conduct the cross-validation at least five-fold, that is, by randomly dividing the data into five sets (ensuring that each compound is well-represented) and repeating the analysis five times. The average lpd over the five models is the resulting summary statistic used to compare models.

3. Results and discussion

The generic Bayesian hierarchical model has been developed for the meta-analysis of human population variability in kinetics. In this section, a reproducible model implementation is discussed, together with applications to derivation of uncertainty factors for subgroups of human populations and inter-phenotypic differences in kinetics for polymorphic metabolism. A case study for the polymorphic CYP2D6 metabolism based on a real data set is presented.

3.1. Model implementation

The model has been implemented in the Stan modelling language [25]. Stan programs are open-source and can be used in various popular statistical software such as R [26].

Stan uses a Hamiltonian Monte Carlo approach to approximate posterior distributions of parameters [27]. In the context of meta-analyses of kinetics data, short Monte Carlo chains may struggle due to data scarcity and statistical identifiability issues discussed above. Therefore, a default recommendation for such model is to use 5,000 iterations as a starting point. Additionally, in some situations (particularly for the modified version of the model), the maximum tree depth parameter in Stan should be set to 15, due to the more complex hierarchy of the model. However, if the value is too low, a warning will be produced automatically. To assess chain convergence the Gelman-Rubin $\hat{R}$ statistic (automatically calculated by Stan) is used [28].

The Stan code for the model is included as part of the Supplementary Materials, together with additional instructions for conducting meta-analyses using the model. Both the generic model and its modification for polymorphic enzymes (using fm and EF parameters) are included in the Supplement. In the latter case, the included

model code allows for sharing parameters across multiple populations. For example, in some situations separate sets of data may be collected for different ethnicities and separate models considered for each. In such cases it is beneficial for the models to share fm parameters, as they are compound-specific and not population-specific. The relevant code to accomplish this is a part of the included code. Calculation of the cross-validation statistic is also included in the "generated quantities" part of the program.

3.2. Applications of the generic model

The generic model can be applied to a range of case studies for the quantification of inter-individual differences between subgroups of human populations and to inform risk assessment with population variability. This includes comparison between neonates, elderly and a reference group (generally healthy adults), as well as the assessment of inter-ethnic and intra-ethnic differences. From the perspective of uncertainty factors, the quantities of interest are the ratio between subgroups R and the subgroup specific variability σ. Figure 2 provides a 3-dimensional plot generalising the relationship between the R, the subgroup-specific variability (characterised by CV, the coefficient of variation; $CV = e^{\sigma} - 1$) and UF's (95th centile) for chemical risk assessment. From the wide range of R values (1-10) and CV (10%-100%), UF (95th centiles) values would range from a minimum of 1 to a maximum of 30. Such a range of parameters and R values would cover most of the range observed in previous meta-analyses of inter-individual kinetic differences in of human populations (Dorne et al, 2005). These meta-analyses showed that human variability in kinetics were mostly below the TK default factor kinetics (3.16) for monomorphic phase I, Phase II enzymes and renal excretion in healthy adults and some instances above for neonates (CYP1A2, glucuronidation, CYP3A4 etc), elderly and PM individuals for polymorphic enzymes (CYP2D6, CYP2C19, NAT-2) (Dorne et al., 2005).

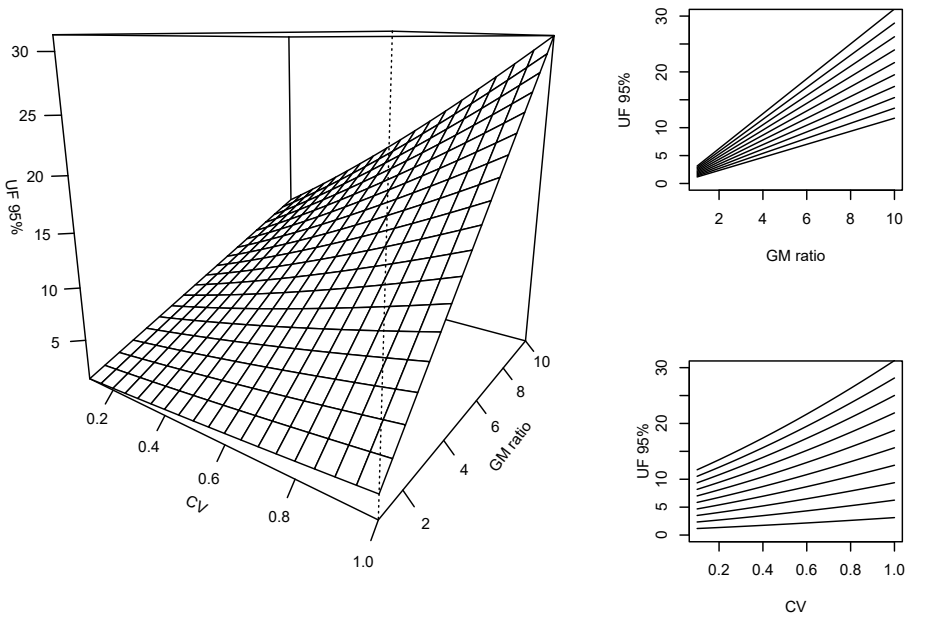


Figure 2. *Human variability in kinetics between subgroups of human population and 95% uncertainty factors for log-normal distributions. Functions of ratios of geometric means (from 1 to 10) and coefficient of variation (from 0.1 to 1) are plotted for the 2-dimensional figures (right-hand panel) as cross-sections through the 3-dimensional plot.*

3.3. Analysing inter-phenotypic differences in kinetics for polymorphic enzyme metabolism: a CYP2D6 case study

As a case study, an analysis of inter-phenotypic differences was performed for CYP2D6 polymorphism and uncertainty factors were derived. An updated database of CYP2D6 kinetic parameters from an earlier work was used to provide realistic input data [29]. The data set consists of 86 study arms reporting *in vivo* estimates of clearance in 8 different compounds: lgm, lv and sample sizes.

Additionally to *in vivo*-specific information, *in vitro* estimates were used to construct prior distributions. Gibbs *et al* [30] applied *in vitro* method to obtain quantitative estimates of fm for many compounds and inferred the hypothetical ratios of CL and AUC in poor to extensive metabolisers. They provide estimates for 13 popular CYP2D6 substrates, with 95% intervals (corresponding to variability in *in vitro* measurements). These reported distributions were used as model priors for the compounds available in the *in vivo* data.

Stan model for analysis of subgroups was used. With the default settings proposed

in the Methods section, the model had no difficulty converging. Figure 3 illustrates the model output, with predictions plotted against the observed data for logarithms of geometric means and variances. Note that in the case of variances the comparison is made between true σ parameters and sample variances.

While the overall statistical fit to data is good, larger residuals are observed in the right tail for the EM polymorphism reflecting large variability in this subgroup and the multi-allelic nature of the CYP2D6 enzyme (from 0 copies to 13 copies). With regards to variance, there is a clear difference between polymorphic phenotypes but limited evidence of differences between compounds. While for readability only the mean values are plotted, validation of the model for variances should also include plotting of sampling distribution intervals or calculating their coverage of reported lv values. In this case study, 7 out of 86 values of lv fall outside of the 95% intervals of sampling distribution given the average estimated σ values and observed sample sizes.

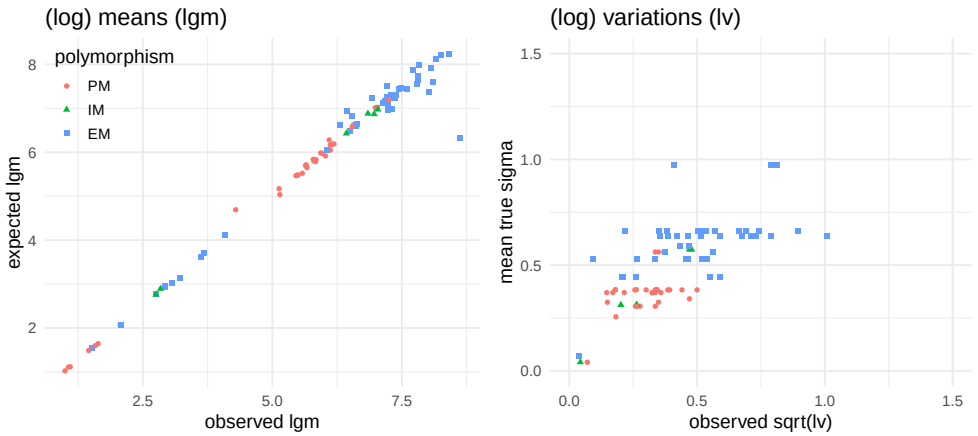


***Figure 3**: Meta-regression for the prediction of inter-phenotypic differences in the kinetics of CYP2D6 substrates. The quantities on y-axis are the modelled means. Bayesian credibility intervals are omitted for readability.*

Example of a graphical validation with regards to the ratios of clearances is shown in Figure 4. In assessing the predicted versus observed ratios (red points in panel C), the variability in the enzyme function (panel A) is combined with the variability in the posterior distributions of fm (panel B). Visually, the 95% intervals of predictive distributions for ratios include all observed ratio values, with exception of tolterodine and propafenone, where only one or two ratios were observed. This suggests that while the model findings can be generalised across compounds, for these two compounds more data should be collected as the model may not be correct.

For some compounds, fm posteriors can be very vague if no prior is fixed (e.g. mexiletine), as shown in the bottom-left panel of Figure 4. This happens when the prior distribution is non-informative and there is no or little data on the inter-phenotypic ratios between kinetic parameters. However, the hierarchical structure of the model is robust to such data gaps and the MCMC procedure had no problems converging.

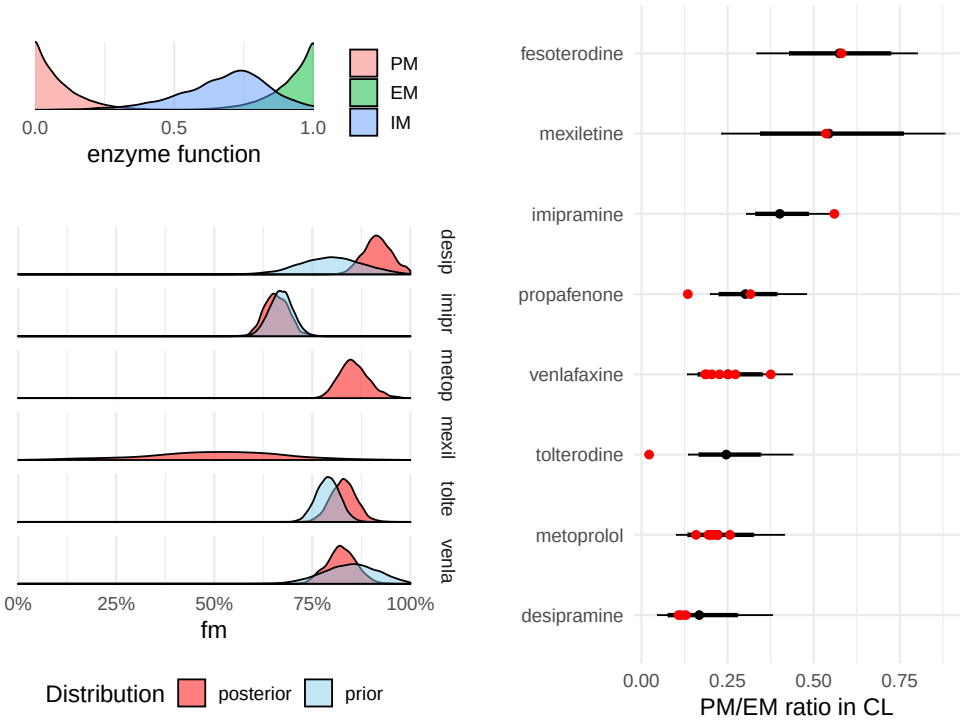


Figure 4*: Analysis of inter-phenotypic kinetic differences for compounds metabolised via the CYP2D6 polymorphic isoform; model validation plots: (A) Posterior predictive distributions for enzyme functioning EF. (B) Comparison of posterior and prior distributions for fraction metabolised. Six representative compounds are shown, with the informative priors based on Table 3 in [30]. (C) Predictive distributions for CL PM/EM ratios of geometric means in CYP2D6, on a substrate-by-substrate basis. Black points are means and bold/thin bars are 80%/95% credibility intervals. Red points are ratios observed in the available studies. Intervals take into account variation in EFs (panel A) as well as in fm (panel B).*

Results of the model are presented in Table 1. In the case of CYP2D6, differences across groups are larger than within groups, as seen on Figure 5. Therefore, there was no need to use the simulation approach to deriving UFs, which are calculated as the ratio of median in EM group to 5th percentile in the PM group. The UFs for each compound are derived and reported as means, with Bayesian 95% uncertainty intervals. Mean UFs range from 3.1 to 12 for the eight compounds, while the geometric mean EM/PM ratios range from 1.7 to 5.9. The UFs are larger than the ratios, due to the within-PM-subgroup variance, ranging from 0.26 to 0.56.

To highlight the role that both ratios and variance estimates play in determining the UFs in the CYP2D6 case study, the estimated mean UFs are also presented in Figure 6, a modification of the 3D plot from Figure 2, with the results of the case study added.

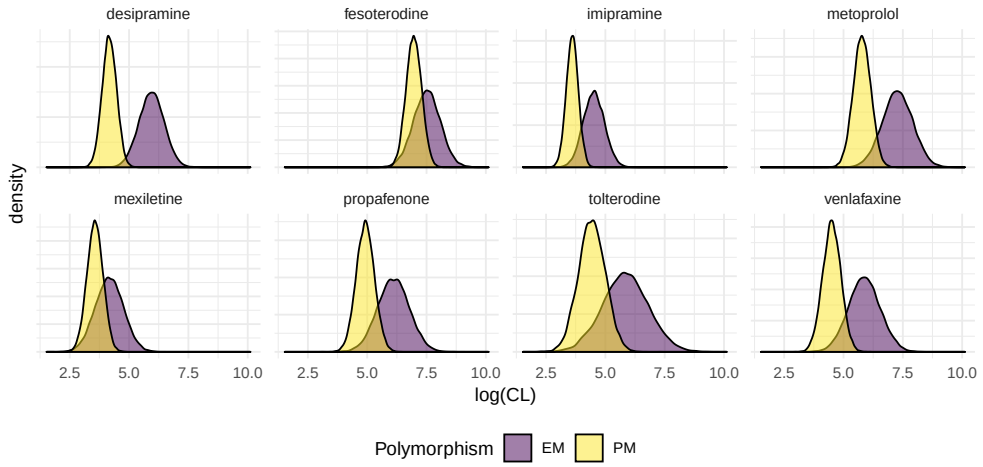


Figure 5: Model-estimated distributions of clearance for extensive (EM) and poor metabolisers (PM), visualised on logarithmic scale.

Compound	EM	PM	PM/EM	EM	PM	UF 95
desipramine	5.9	4.2	0.17	0.53	0.31	12 (6.1, 18.2)
fesoterodine	7.5	7.0	0.58	0.56	0.32	3.1 (2.3, 4)
imipramine	4.5	3.6	0.40	0.44	0.26	3.9 (3.1, 4.6)
metoprolol	7.3	5.8	0.22	0.64	0.37	9 (5.8, 12.2)
mexiletine	4.2	3.6	0.55	0.59	0.34	3.6 (2.3, 5.2)
propafenone	6.1	4.9	0.30	0.66	0.38	6.5 (4.8, 8.2)
tolterodine	5.8	4.4	0.25	0.97	0.56	11.1 (7.2, 15.3)
venlafaxine	5.9	4.5	0.25	0.66	0.38	7.9 (5.6, 10.4)

Table 1: Results of the CYP2D6 case study. For each drug mean clearance and its variability are quantified both in extensive and poor metabolisers. 95% uncertainty factor depends on both ratio of geometric means and coefficient of variation and is provided together with Bayesian 95% uncertainty interval.

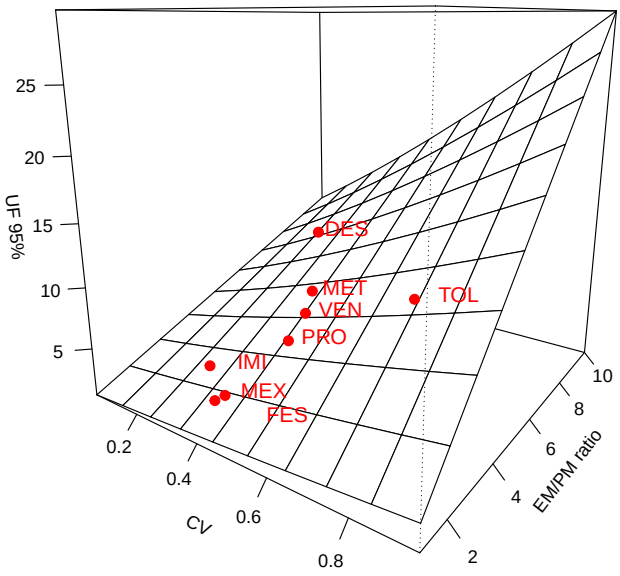


Figure 6: Model-estimated mean 95% uncertainty factors for clearance parameters in nine compounds metabolised by CYP2D6. UFs are a function of ratio of means between extensive and poor metabolisers and the estimated population variability.

4. Conclusion and future directions

A generic Bayesian hierarchical model for the meta-analysis of kinetic data has been developed to quantify population variability in subgroups of human populations and to derive UFs for chemical risk assessment. Applications to chemical risk assessment have been illustrated: one, characterisation of inter-individual differences for subgroups of human populations and two, quantification of inter-phenotypic differences in CYP2D6 metabolism.

The hierarchical structure of the model presented addresses a number of data features that are common in evidence synthesis for chemical risk assessment, namely: 1) low sample sizes, 2) a hierarchical structure with subgroup-specific, compound-specific, study-specific and parameter-specific data, 3) potentially large heterogeneity across studies, and 4) data gaps.

There is a number of advantages to using the generic approach presented here over standard meta-analysis models, such as inverse variance weighting method or other fixed effect models of meta-analysis. The Bayesian approach avoids problems with estimating variance that can occur in some maximum likelihood approaches to mixed modelling [31,32]. Treating the true study variance as an unknown parameter both improves estimation for the means and allows modelling of effect of a compound

or subgroup on variance. This is in contrast to standard meta-analysis approaches where only the behaviour of means is considered. Accounting for means and variances simultaneously is essential to correctly estimate UFs, as shown in Figures 2 and 6.

Applying the model to polymorphic enzymes offers further benefits over the generic approach. The model allows for capturing variability in ratios through integration of compound-specific fm's and population-specific EF's. Both parameters are interpretable on a biological basis. In addition, informative priors for both fm and EF can often be collected from the literature and incorporated in the model. Moreover, the presented model can be generalised across many parameters and populations by using the code included in the Supplementary Materials.

However, the feasibility of modelling population variability in kinetics will depend on availability of the relevant data. Data gaps may lead to high uncertainty in such results. These gaps and their impact on the outcome of the meta-analysis should be described transparently in each case. The assumption of ratios being variable across studies and that the reported variance is not equal to true variance are justified by variability in ratios and sample sizes, as typically reported in the literature. However, this study does not quantify the benefit of the proposed approach to variance modelling compared to standard meta-analysis approaches.

With such limitations in mind, further work on theoretical aspects of the model may include: 1) investigating the joint analysis of multiple kinetic parameters, leading to "borrowing of strength" across parameters [33]; 2) quantifying the impact of small sample sizes and use of different measures of dispersion on the results (possibly through a simulation study, similarly to other published work [34]); 3) allowing for modelling of measurement errors related to different measures of dispersion, including standard deviations, standard error of the mean and confidence intervals.

This Bayesian hierarchical model has also been applied to investigate metabolic interactions. Quignot et al ([35]) presented a meta-analysis of kinetic data in humans following inhibition of CYP3A4 metabolism with grapefruit juice and CYP3A4 induction with St John's wort. The approach is similar to the above described for comparing kinetics in CYP2D6 phenotypes since fm is taken into account together with bioavailability, while the reference group is the control and the subgroups are the grapefruit juice or St John's wort treatments. Uncertainty factors were derived based on the coefficients of variation as well as the impact of the interaction on the geometric means ratio. In the case of no interaction, the variability was found to be within the default TK UF of 3.16, whereas the estimated UF values were up to 6 times higher than the default UF for maximum inhibition by grapefruit juice. For St John's wort, the maximum induction resulted in UFs twice the value of the default TK UF.

From an applied perspective, broadening the application of this generic model could be beneficial to the development of non-testing methods for the use of New Approach Methods in chemical risk assessment including:

1. Quantifying human variability for major human metabolic and excretion path-

ways including Phase I, phase II metabolism, renal excretion and transporters. This application is currently being investigated as a collaborative project between the European Food Safety Authority (EFSA), national agencies (ANSES, ISS) and academia (University of Utrecht). A first application for the analysis of human variability in CYP3A4 metabolism is published in this special issue [36]. This also include other polymorphic enzymes such as CYP2C9, CYP2C19 as well as phase II enzymes (N-actylation (NAT-2) with fast and slow acetylators (FA and SA), glutathione-s-transferases and UDP-glucuronosyltransferases) and transporters (Organic anion Transporter Proteins (OATP) and P-glycoproteins) [11,37].

2. Predicting human kinetic parameters (e.g Cmax, half-life, clearance) using quantitative *in vitro-in vivo* extrapolations (QIVIVE) from the integration of human *in vitro* metabolism data and metabolic variability distributions into human toxicokinetic (TK) or physiologically-based toxicokinetic (PB-TK) models for chemicals of relevance to food and feed safety (e.g. pesticides, food and feed additives, contaminants, etc.).
3. Integration of variability in metabolism into TK-toxicodynamic (TK-TD) or PB-TK-TD models to quantify the impact of metabolism or variability in TK (e.g. subgroups of human populations, polymorphisms) on toxicological outcomes. For example, a TK model or PB-TK model would quantify internal concentrations of a chemical or a metabolite in the target organ; for the TD part, dose-response modelling would be performed using benchmark dose modelling on an internal dose basis.
4. Development of quantitative *in vitro in vivo* extrapolations models (QIVIVE) integrating isoform specific *in vitro* information and kinetic variability distributions for different subgroups and phenotypes of human populations as a basis for refining the derivation of reference points/points of departure in risk assessment based on internal dose and for the refinement of UFs such as pathway-derived UFs and chemical-specific adjustment factors.
5. Applications of the Bayesian hierarchical model to the meta-analysis of TK or toxicity data in species of relevance to ecological risk assessment (e.g. daphnia, fish, earth worms). These applications can provide tools to assess interspecies differences in kinetics or toxicity for specific substances or group of substances or support the development of internal species-sensitivity distributions [38].

Acknowledgements

The discussed model has been developed at Certara during a project financed by the European Food Safety Authority (EFSA) under contract CFT/EFSA/EMRISK/2012/01. The authors wish to thank Prof. Marc Aerts from University of Hasselt for his valuable comments on the drafts of this paper.

References

[1] W. (World Health Organization), Food Safety Project to update the principles and methods for the assessment of chemicals in food Principles and methods for the risk assessment of chemicals in food, 2009.

[2] SCENIHR, Memorandum on the use of the scientific literature for human health risk assessment purposes weighing of evidence and expression of uncertainty, (2012).

[3] A. Hardy, D. Benford, T. Halldorsson, M.J. Jeger, H.K. Knutsen, S. More, H. Naegeli, H. Noteborn, C. Ockleford, A. Ricci, G. Rychen, J.R. Schlatter, V. Silano, R. Solecki, D. Turck, E. Benfenati, Q.M. Chaudhry, P. Craig, G. Frampton, M. Greiner, et al., Guidance on the use of the weight of evidence approach in scientific assessments, EFSA Journal, 15 (2017) e04971.

[4] R. Truhaut, The concept of the acceptable daily intake: An historical review, Food Additives and Contaminants, 8 (n.d.) 151–162.

[5] A.G. Renwick, Data-derived safety factors for the evaluation of food additives and environmental contaminants, Food Additives & Contaminants, 10 (1993) 275–305.

[6] A.G. Renwick, N.R. Lazarus, Human Variability and Noncancer Risk Assessment An Analysis of the Default Uncertainty Factor, Regulatory Toxicology and Pharmacology, 27 (1998) 3–20.

[7] B.D. Naumann, K.C. Silverman, R. Dixit, E.C. Faria, E.V. Sargent, Case Studies of Categorical Data-Derived Adjustment Factors, Human and Ecological Risk Assessment: An International Journal, 7 (2001) 61–105.

[8] J.L.C.M. Dorne, K. Walton, A.G. Renwick, Human variability in xenobiotic metabolism and pathway-related uncertainty factors for chemical risk assessment: A review, Food and Chemical Toxicology, 43 (2005) 203–216.

[9] G. Ginsberg, D. Hattis, B. Sonawane, A. Russ, P. Banati, M. Kozlak, S. Smolenski, R. Goble, Evaluation of Child/Adult Pharmacokinetic Differences from a Database Derived from the Therapeutic Drug Literature, Toxicological Sciences, 66 (2002) 185–200.

[10] J.L.C.M. Dorne, K. Walton, A.G. Renwick, Human variability in glucuronidation in relation to uncertainty factors for risk assessment, Food and Chemical Toxicology, 39 (2001) 1153–1173.

[11] J.L.C.M. Dorne, K. Walton, A.G. Renwick, Polymorphic CYP2C19 and N-acetylation: Human variability in kinetics and pathway-related uncertainty factors, Food and Chemical Toxicology, 41 (2003) 225–245.

[12] J.L.C.M. Dorne, K. Walton, A.G. Renwick, Human variability for metabolic pathways with limited data (CYP2A6, CYP2C9, CYP2E1, ADH, esterases, glycine and sulphate conjugation), Food and Chemical Toxicology, 42 (2004) 397–421.

[13] K. Walton, J.L. Dorne, A.G. Renwick, Uncertainty factors for chemical risk assessment: Interspecies differences in the in vivo pharmacokinetics and metabolism of human CYP1A2 substrates, Food and Chemical Toxicology, 39 (2001) 667–680.

[14] K. Walton, J.L. Dorne, A.G. Renwick, Uncertainty factors for chemical risk assessment: Interspecies differences in glucuronidation, Food and Chemical Toxicology, 39 (2001) 1175–1190.

[15] J.L.C.M. Dorne, K. Walton, W. Slob, A.G. Renwick, Human variability in polymorphic CYP2D6 metabolism: Is the kinetic default uncertainty factor adequate?, Food and Chemical Toxicology, 40 (2002) 1633–1656.

[16] A.J. Sutton, J.P.T. Higgins, Recent developments in meta-analysis, Statistics in Medicine, 27 (2008) 625–650.

[17] C. Rigaux, J.-B. Denis, I. Albert, F. Carlin, A meta-analysis accounting for sources of variability to estimate heat resistance reference parameters of bacteria using hierarchical Bayesian modeling: Estimation of D at 1211 C and pH 7, zT and zpH of Geobacillus stearothermophilus, International Journal of Food Microbiology, 161 (2013) 112–120.

[18] W. (World Health Organization), Guidance document on evaluating and expressing uncertainty in hazard characterization 2nd edition, World Health Organization, Geneva, 2018.

[19] M.N. Pieters, H.J. Kramer, W. Slob, Evaluation of the uncertainty factor for subchronic-to-chronic extrapolation: Statistical analysis of toxicity data, Regulatory Toxicology and Pharmacology: RTP, 27 (1998) 108–111.

[20] B.E. Saltzman, Health Risk Assessment of Fluctuating Concentrations Using Lognormal Models, Journal of the Air & Waste Management Association, 47 (1997) 1152–1160.

[21] W. Slob, Strategies in applying statistics in ecological research PhD thesis, Free University Press, Amsterdam, 1986.

[22] S.P. Hozo, B. Djulbegovic, I. Hozo, Estimating the mean and variance from the median, range, and the size of a sample, BMC Medical Research Methodology, 5 (2005) 13.

[23] A. Gelman, Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper), Bayesian Analysis, 1 (2006) 515–534.

[24] J.L.C.M. Dorne, Metabolism, variability and risk assessment, Toxicology, 268 (2010) 156–164.

[25] S.D. Team, The Stan Core Library, (2018).

[26] R.D.C. Team, R: A Language and Environment for Statistical Computing, (2008).

[27] R.M. Neal, MCMC using Hamiltonian dynamics, arXiv:1206.1901 [Physics, Stat], (2012).

[28] A. Gelman, D.B. Rubin, Inference from Iterative Simulation Using Multiple Sequences, Statistical Science, 7 (1992) 457–472.

[29] C. Béchaux, B. Amzal, A. Crépet, J.-L. Dorne, Meta-analysis to better integrate human variability in toxicokinetic: CYP2D6-related uncertainty factors, 2015.

[30] J.P. Gibbs, R. Hyland, K. Youdim, Minimizing Polymorphic Metabolism in Drug Discovery: Evaluation of the Utility of in Vitro Methods for Predicting Pharmacokinetic Consequences Associated with CYP2D6 Metabolism, Drug Metabolism and Disposition, 34 (2006) 1516–1522.

[31] B.M. Bolker, M.E. Brooks, C.J. Clark, S.W. Geange, J.R. Poulsen, M.H.H. Stevens, J.-S.S. White, Generalized linear mixed models: A practical guide for ecology and evolution, Trends in Ecology & Evolution, 24 (2009) 127–135.

[32] X.A. Harrison, L. Donaldson, M.E. Correa-Cano, J. Evans, D.N. Fisher, C.E.D. Goodwin, B.S. Robinson, D.J. Hodgson, R. Inger, A brief introduction to mixed effects modelling and multi-model inference in ecology, PeerJ, 6 (2018).

[33] R.D. Riley, Multivariate meta-analysis: The effect of ignoring within-study correlation, Journal of the Royal Statistical Society: Series A (Statistics in Society), 172 (2009) 789–811.

[34] L. Lin, Bias caused by sampling error in meta-analysis with small sample sizes, PLOS ONE, 13 (n.d.) e0204056.

[35] N. Quignot, W. Wiecek, B. Amzal, J.-L. Dorne, The YinYang of CYP3A4: A Bayesian meta-analysis to quantify inhibition and induction of CYP3A4 metabolism in humans and refine uncertainty factors for mixture risk assessment, Archives of Toxicology, (2018).

[36] K. Darney, E. Testai, E. Buratti, N. Kramer, E. Kasteel, E. Di Consiglio, C. Bechaux, J.L. Dorne, Inter- and intra-ethnic differences in CYP3A4 metabolism: A Bayesian meta-analysis for the refinement of uncertainty factors in chemical risk assessment, (2019).

[37] L.-A. Clerbaux, A. Paini, A. Lumen, H. Osman-Ponchet, A.P. Worth, O. Fardel, Membrane transporter data to support kinetically-informed chemical risk assessment using non-animal methods: Scientific and regulatory perspectives, Environment International, 126 (2019) 659–671.

[38] J.-H. Kwon, S.-Y. Lee, H.-J. Kang, P. Mayer, B.I. Escher, Including Bioconcentration Kinetics for the Prioritization and Interpretation of Regulatory Aquatic Toxicity Tests of Highly Hydrophobic Chemicals, Environmental Science & Technology, 50 (2016) 12004–12011.

Code for meta-analysis models in Stan

This supplement includes generic model codes. Readers should refer to main text of the paper to understand the notation.

Conducting inference using Stan – recommendations

In the context of meta-analyses of kinetics data, short Monte Carlo chains may struggle due to data scarcity and statistical identifiability issues we discussed in the paper. Therefore a default recommendation for such model is to use 5,000 iterations as a starting point. Additionally, for more complicated models (such as the one presented in Section 3.2) the maximum tree depth parameter in Stan should be set to 15, due to more complex hierarchy of the model. (If it is too low, a warning will automatically be produced.) To assess chain convergence the Gelman-Rubin $\hat{R}$ statistic (calculated automatically by Stan) is used.

Divergent transitions may be reported by Stan, especially when input data are sparse. This can occur even when the Rhat statistic is low (below 1.01). In such cases additional diagnostics should be produced to ensure convergence to the target distribution. As a basic step adapt_delta option should be increased, per automated advice of the tool and pairs() plot should be produced for sigma and gamma parameters (see code below) to assess if any divergences occur in a systematic way, especially for particular compounds or studies. Users should refer to the manual for more detailed advice for how to examine the divergent transitions (https://mc-stan.org/misc/warnings.html).

Generic model

This model is for random effects on ratios, as presented in Section 2.1. The model can be changed to fixed effects on ratios by removing some comments and commenting out lines after `random-effect` comments.

```
/*
  Stan model for modelling ratios of subgroups
*/

data {
  //counts
  int  n;
  int  n_drugs;
  int  n_studies;
  int  n_groups;

  //identifiers d(i), s(i), g(i)
  int  drug[n];
  int  study[n];
  int  group[n]; //code reference group as 1

  //observed data
  real ss[n];
  real lgm[n];
  real<lower=0> lv[n];
}

parameters {
  real mu_drug[n_drugs];
  real mu_study[n_studies];
  real<lower=0> sigma_study;
  real gamma_group[n_groups-1];
  real gamma_drug[n_drugs];

  // random effect of group:
  // real logratio[n];
  // real mu_group[n_groups - 1];
  // real<lower=0> sigma_group;

  // fixed effect of group:
  real logratio[n_groups-1];
}

transformed parameters { //means and SDs in each study arm
```

```
  real sigma[n];
  real mu[n];
  for(i in 1:n){
    if(group[i] != 1){
      // random effect of group
      // mu[i] = mu_drug[drug[i]] + mu_study[study[i]] + logratio[n];
      // fixed effect of group
      mu[i] = mu_drug[drug[i]] + mu_study[study[i]] +
              logratio[group[i] - 1];
      sigma[i] = exp(gamma_drug[drug[i]] + gamma_group[group[i] - 1]);
    }else{
      mu[i] = mu_drug[drug[i]] + mu_study[study[i]];
      sigma[i] = exp(gamma_drug[drug[i]]);
    }
  }
}

model {
  for(i in 1:n_studies)
    mu_study[i] ~ normal(0, sigma_study);
  sigma_study   ~ normal(0, 5);
  mu_drug       ~ normal(0, 10);
  gamma_drug    ~ normal(0, 5);
  gamma_group   ~ normal(0, 5);

  // random effect of group:
  // mu_group ~ normal(0, 10);
  // sigma_group ~ normal(0, 2.5);

  // fixed effect of group:
  logratio ~ normal(0, 2.5);

  for(i in 1:n){
    // random effect of group:
    // if(group[i] == 1)
    //   logratio[i] ~ normal(0, .0001);
    // else
    //   logratio[i] ~ normal(mu_group[group[i] - 1], sigma_group);
    //observed quantities:
    lgm[i] ~ normal(mu[i], sigma[i]/sqrt(ss[i]));
    //chi-sq is Gamma(njk - 1 / 2, tau_j*(n_jk-1)/2)
    lv[i]  ~ gamma((ss[i] - 1) / 2, (ss[i]-1)/(2*(sigma[i]^2)));
  }
}
```

Model for polymorphisms

```
/*
  Stan model for modelling polymorphisms of kinetic parameters
*/

data {
  //counts
  int  n;
  int  n_drugs;
  int  n_studies;
  int  n_polymorphisms;
  //identifiers
  int  drug[n];
  int  study[n];
  int  polymorphism[n]; //code the polymorphism with EF=0 as 1,
                        //and  the one with EF=1 as n_polymorphisms

  real<lower=0> fm_mean_prior[n_drugs];

  //observed data on means & variation:
  real ss[n];
  real lgm[n];
  real<lower=0> lv[n];
}

parameters {
  real mu_drug[n_drugs];
  real mu_study[n_studies];
  real<lower=0> sigma_study;
  real gamma_group[n_polymorphisms-1];
  real gamma_drug[n_drugs];

  // random effect of group:
  real<lower=0, upper=1> mean_ef[n_polymorphisms-2];
  real<lower=0, upper=1> fm[n_drugs];
  real<lower=0, upper =1> ef[n];
  real<lower=0> sigma_ef;
}

transformed parameters {
  real mu[n];
  real sigma[n];
  for(i in 1:n)
    //this is for CL
```

```
    mu[i] = mu_drug[drug[i]] + mu_study[study[i]] +
            log(ef[i]*fm[drug[i]] + 1-fm[drug[i]]);
    //for AUC use change the last "+"" to "-""
    //because R_AUC = 1/R_CL
  for(i in 1:n){
    sigma[i] = exp(gamma_drug[drug[i]]);
    if(polymorphism[i] > 1)
      sigma[i] = sigma[i]*exp(gamma_group[polymorphism[i]-1]);
  }
}

model {
  //priors dealing with means (mu):
  mu_drug  ~ normal(0, 10);
  mu_study ~ normal(0, sigma_study);
  sigma_study ~ normal(0, 5);

  //priors for fm:
  for(i in 1:n_drugs)
    fm[i] ~ normal(fm_mean_prior[i], .1);
  sigma_ef ~ normal(0, 1);

  //priors dealing with variances (gamma):
  gamma_drug ~ normal(0, 5);
  gamma_group ~ normal(0, 5);
  for(i in 1:(n_polymorphisms-2))
    mean_ef[i] ~ uniform(0,1);

  for(i in 1:n) {
    //enzyme function:
    if(polymorphism[i] == 1)
      ef[i] ~ normal(0, sigma_ef);
    if(polymorphism[i] == n_polymorphisms)
      ef[i] ~ normal(1, sigma_ef);
    if((polymorphism[i] > 1) && (polymorphism[i] < n_polymorphisms))
      ef[i] ~ normal(mean_ef[polymorphism[i]-1], sigma_ef);

    //observed quantities:
    lgm[i] ~ normal(mu[i], sigma[i]/sqrt(ss[i]));
    //chi-sq is Gamma(njk - 1 / 2, tau_j*(n_jk-1)/2)
    lv[i]  ~ gamma((ss[i] - 1) / 2, (ss[i]-1)/(2*(sigma[i]^2)));
  }
}
```

Archives of Toxicology
https://doi.org/10.1007/s00204-018-2325-6

TOXICOGENOMICS

The Yin–Yang of CYP3A4: a Bayesian meta-analysis to quantify inhibition and induction of CYP3A4 metabolism in humans and refine uncertainty factors for mixture risk assessment

Nadia Quignot[1] · Witold Wiecek[2] · Billy Amzal[1] · Jean-Lou Dorne[3]

Received: 1 August 2018 / Accepted: 2 October 2018

Abstract
Quantifying differences in pharmacokinetics (PK) and toxicokinetics (TK) provides a science-based approach to refine uncertainty factors (UFs) for chemical risk assessment. Cytochrome P450 (CYP) 3A4—the major hepatic and intestinal human CYP—and the P-glycoprotein (Pgp) transporter share a vast range of common substrates for which PK may be modulated through inhibition or induction in the presence of grapefruit juice (GFJ) or St. John's wort (SJW), respectively. Here, an extensive literature search was performed on PK interactions for CYP3A4 and Pgp substrates after oral co-exposure to GFJ and SJW. Relevant data from 109 publications, extracted for both markers of acute (C_{max}) and chronic [clearance and area under the plasma concentration–time curve (AUC)] exposure, were computed into a Bayesian hierarchical meta-analysis model. Bioavailability (F) and substrate fraction metabolised by CYP3A4 (Fm) were identified as the variables exhibiting the highest impact on the magnitude of interaction. The Bayesian meta-regression model developed provided good predictions for magnitudes of inhibition (maximum 5.3-fold with GFJ) and induction (maximum 2.3-fold with SJW). Integration of CYP3A4 variability, F, Fm and magnitude of interaction provided the basis to derive a range of CYP3A4 and Pgp-related UFs. Such CYP3A4 and Pgp-related UFs can be derived in the absence of human data using in vitro TK evidence for CYP3A4/Pgp inhibition or induction as conservative in silico options. The future development of quantitative in vitro–in vivo extrapolation models for mixture risk assessment is discussed with particular attention to integrating human in vitro and in vivo P/TK data on interactions with pathway-related variability.

Keywords CYP3A4 · Interindividual variability · Kinetic interactions · Mixtures · Risk assessment · Uncertainty factors

Introduction

Human variability in pharmacokinetics (PK) or toxicokinetics (TK), expressed as interindividual variability, may impact drug efficacy (pharmacodynamics, PD) or chemical toxicity (toxicodynamics, TD). In a similar manner, interactions between chemicals may affect P/TK, P/TD or both dimensions. Typical examples of such interactions at the kinetic level include interactions between drugs (drug–drug interactions, DDI), food and drugs (FDI), herbs and drugs (HDI) (Fujita 2004), or chemical–chemical interactions (CCI).

A number of inhibitors and inducers are known to impact the P/TK of compounds through interaction with the major metabolising enzyme cytochrome P450 3A4 (CYP3A4) isoform and/or the major efflux intestinal transporter P-glycoprotein (Pgp), which have considerable substrate affinity overlap (Staud et al. 2010). These compounds affect P/TK by acting on both oral bioavailability [including CYP3A4-mediated first-pass metabolism in intestine and liver (Almazroo et al. 2017) and Pgp-mediated absorption (Ambudkar et al. 1999)], and systemic liver metabolism [through CYP3A4 enzyme (Zhou 2008)]. The P/TK consequences are reflected through modifications of both acute exposure: blood maximum

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00204-018-2325-6) contains supplementary material, which is available to authorized users.

✉ Nadia Quignot
n.quignot@analytica-laser.com

[1] Analytica LASER, Paris, France

[2] Analytica LASER, London, UK

[3] European Food Safety Authority, Parma, Italy

Published online: 08 October 2018

concentration (C_{max}) and chronic exposure: area under the plasma concentration–time curve (AUC), clearance (CL), and half-life ($t_{1/2}$).

Since DDI, FDI, HDI and CCI are a public health issue, many research efforts have led to the development of evidence-based models to predict magnitudes of interaction in the presence of inhibitors or inducers for a range of enzymes, to adjust the therapeutic doses plan, to minimise the risk of adverse health effects and maximise benefits of the therapeutic compounds (Choi and Ko 2017; Hennessy et al. 2016; Won et al. 2012). These include predictive models based on in silico analysis, in vitro or in vivo data (Einolf 2007; Roy and Roy 2009) and physiologically based PK (PBPK) models based on in vivo and in vitro data (EMA 2012; FDA 2012; Jamei 2016). A critical step in the development of such models using in silico and in vitro data is the inclusion of in vivo data to update the model to refine predictions and support its performance (Zhuang and Lu 2016).

One of the most well-known inhibitors of CYP3A4 and Pgp is grapefruit juice (GFJ) (Seden et al. 2010); its counterpart as an inducer is St. John's wort (SJW), a plant used as an anti-depressant herbal remedy (Rahimi and Abdollahi 2012). GFJ and SJW interactions with a large number of therapeutic drugs are well described in the literature (Bailey and Dresser 2004; Isoherranen et al. 2004; Kober et al. 2008). Nevertheless, magnitude of interactions between CYP3A4/Pgp substrates and inhibitors or inducers may be not sufficiently quantified to adequately predict the impact on the established benefit/risk profile (Won et al. 2012). This is due to interindividual variability, specificities of substrates and inhibitors/inducers, as well as differences in methodologies between clinical studies.

In human risk assessment, addressing interspecies differences and human variability for hazard characterisation has been traditionally performed using the default uncertainty factor (UF) of 100-fold (10×10) applied to chronic or sub-chronic toxicity data to derive safe levels of exposure. Recent guidance documents have proposed to replace default UFs with chemical-specific adjustment factors (CSAFs), or more generally UFs derived using relevant in vivo P/TK and/or P/TD data to reduce uncertainty or better characterise variability (Bhat et al. 2017). In this context, CYP3A4 pathway-specific UFs have been derived using human PK data for CYP3A4 substrates to account for interindividual variability in the isoform and compared with the default TK UF of 3.16 (Dorne et al. 2003a). Such pathway-related UFs constitute an intermediate option between default values and CSAFs and can be applied when human in vitro metabolism data are available for specific isoforms but no in vivo data are available. In this study, the current 3.16 kinetic default factor was shown to cover at least 99% of healthy adults. However, sources of variability including overlapping substrate specificities between CYP3A4 and Pgp, as well as impact of inhibitors and inducers on PK parameters of probe substrates, were not assessed.

In a mixture risk assessment context, the impact of inhibitors and inducers on metabolism and TK and the associated consequences on the derivation of UFs need to be addressed using a data-based approach and constitute the aim of this study. Here, an extensive literature search was conducted and a database of published PK studies was constructed. A meta-analysis of PK data for parameters reflecting acute exposure (C_{max}) and chronic exposure (AUC and CL) in the presence of GFJ or SJW was performed. Magnitudes of interaction were derived and correlations between such magnitudes of interaction and substrate-specific parameters were fitted. A meta-regression model integrating the human PK data and the fitted correlations was developed to predict the magnitude of interaction and associated variability based on substrate characteristics. CYP3A4-related UFs were derived for single compounds and binary mixtures taking into account CYP3A4 inhibition and induction.

Materials and methods

Data collection

Data on PK interactions between substrates of CYP3A4 and/or Pgp and GFJ or SJW were collected through an Extensive literature search (ELS) (EFSA 2010; FDA 2009). ELS was performed as published by Quignot et al. (2015) using online databases PubMed (http://www.ncbi.nlm.nih.gov/pubmed), Embase® (http://www.embase.com), Cochrane (all databases, http://www.cochrane.org), and Web of Science™ (http://www.webofknowledge.com) and covering literature up to February 2018. An initial screening of titles and abstracts, followed by a second screening based on full texts, were performed against the inclusion criteria (healthy adults, oral route of exposure, quantitative data including statistical descriptors about PK parameters C_{max}, AUC and CL), to identify relevant peer reviewed publications. The following exclusion criteria were applied: to quantify variability due to CYP3A4 metabolism, delayed or slow release formulations were excluded to screen out variability in absorption as a co-variate; studies in which humans were exposed to more than two compounds (including GFJ and SJW) were excluded; articles published in a language other than English were excluded. Substrates of CYP3A4 and Pgp were identified from the literature. For each compound, the fraction metabolised by CYP3A4 (Fm) was determined. For substrates concluded as "major" substrates based on in vitro data but without data on Fm, a default Fm of 0.8 was applied. For substrates concluded as "minor" substrates, a default Fm of 0.2 was assigned. Single substrate oral bioavailability (F) was also extracted from the literature. When bioavailability was described with an interval, the mean

was calculated. Substrate-specific F and CYP3A4 Fm are presented in Supplementary material S1. PK interaction data (binary mixtures) for CYP3A4 and Pgp substrates for GFJ and SJW were computed and analysed for markers of acute and chronic exposure as detailed below.

Data analysis

Data harmonisation

Markers of acute (C_{max}) and chronic (CL and AUC) oral exposure were collected for each individual study and data were harmonised for the purpose of the meta-analyses:

- C_{max} values were normalised to ng/ml and adjusted for dose and body weight (mg/kg body weight);
- AUC values were normalised to ng/ml h and adjusted for dose and body weight (mg/kg body weight);
- CL values were adjusted to body weight (when not reported as such in the study) and normalised to be expressed in ml/min/kg.

When body weights were not reported, the following mean adult body weights were used: 70 kg (males), 60 kg (females) or 65 kg (mixed males and females).

Data in the kinetic studies were most of the time either reported as arithmetic means (X) and standard deviations (SD) or as geometric means (GM) and geometric standard deviations (GSD). Since kinetic data are generally recognised to be lognormally distributed (Dorne et al. 2001a; Naumann et al. 1997; Renwick and Lazarus 1998) and that the geometric mean (GM) and geometric standard deviation (GSD) are better appropriate to summarise a lognormal distribution, all the kinetic data in this work are described as GM and GSD. When these measures were not reported, they were estimated for each individual study using the following equations:

$$\mathrm{GM} = \frac{X}{\sqrt{(1+\mathrm{CV_N^2})}}, \tag{1}$$

$$\mathrm{GSD} = \exp\left(\sqrt{\ln\left(1+\mathrm{CV_N^2}\right)}\right), \tag{2}$$

where $\mathrm{CV_N}$ is the coefficient of variation given by:

$$\mathrm{CV_N} = \frac{\mathrm{SD}}{X}. \tag{3}$$

In some studies, SD was not reported and was estimated from the standard error SE (SEM), $\mathrm{CV_N}$ and the 95% confidence interval of the mean with the following equations:

$$\mathrm{SD} = \sqrt{n} \times \mathrm{SE}, \tag{4}$$

$$\mathrm{SD} = \mathrm{CV_N} \times X, \tag{5}$$

$$\mathrm{SD} = \frac{\mathrm{UCI}-\mathrm{LCI}}{2t_{0.975,n-1}} \times \sqrt{n}, \tag{6}$$

where n is the sample size, UCI and LCI refer to the upper and lower bounds of confidence interval and $t_{0.975,n-1}$ is the 97.5th percentile of the t distribution with $t_{0.975,n-1}$ degrees of freedom.

The coefficient of variation for lognormally distributed data is given by:

$$\mathrm{CV_{LN}} = \sqrt{\exp\{[\ln(\mathrm{GSD})]^2\} - 1}. \tag{7}$$

Magnitude of effects and statistical significance of individual ratios

Magnitudes of the PK interactions for each study were expressed as the ratios between geometric means of parameters for the single compound and the binary mixtures. When stated, the ratio was expressed on a harmonised scale starting at 1 to reflect changes in internal dose for either inhibition or induction. These ratios were calculated for each kinetic parameter (C_{max}, AUC and CL).

$$\mathrm{GM\ ratio} = \frac{\mathrm{GM_I}}{\mathrm{GM_C}}, \tag{8}$$

where $\mathrm{GM_I}$ refers to geometric mean (lognormal distribution) for interaction data and $\mathrm{GM_C}$ refers to geometric mean (lognormal distribution) for control data.

Confidence intervals for GM ratio, with UCI and LCI referring to upper and lower bounds, were calculated as follows:

$$\mathrm{LCI} = \exp\left(\ln(\mathrm{GM\ ratio}) - 1.96\ \mathrm{SE}_{\ln(\mathrm{GM\ ratio})}\right), \tag{9}$$

$$\mathrm{UCI} = \exp\left(\ln(\mathrm{GM\ ratio}) + 1.96\ \mathrm{SE}_{\ln(\mathrm{GM\ ratio})}\right), \tag{10}$$

$$\mathrm{SE}_{\ln(\mathrm{GM\ ratio})} = \sqrt{\frac{\ln^2(\mathrm{GSD_C}) + \ln^2(\mathrm{GSD_I})}{\min(n_C, n_I)}}, \tag{11}$$

where C refers to control, I to interaction, and weight n to the minimum sample size between control and interaction data.

For PK parameters, when statistical tests were not conducted, not described, or providing different results in the literature, statistical significance was tested for each study using t statistic. Given the geometric mean of the control data $\mathrm{GM_C}$, the geometric mean of the interaction data $\mathrm{GM_I}$, the geometric standard deviation of the control data GSD and the sample size n, the t statistic was calculated as:

$$t = \frac{\log(\mathrm{GM_C}) - \log(\mathrm{GM_I})}{\log(\mathrm{GSD}) \Big/ \sqrt{n}}, \tag{12}$$

which under the null hypothesis, follows a Student distribution. Interaction ratios related to a p value < 0.10 were considered as statistically significant, accounting for calculation biases.

Relationship between substrate-specific parameters and magnitudes of interaction

Correlation between substrate-specific parameters (i.e., oral bioavailability F and fraction metabolised by CYP3A4 Fm) and magnitudes of interaction (i.e., GM ratio) was assessed.

$$\mu_i = \begin{cases} \mu_{\text{drug}(i)} + \mu_{\text{study}(i)} + \gamma 1_{\text{int}(i)} F_i + \gamma 2_{\text{int}(i)} \text{Fm}_i, & \text{if int}(i) = \text{GFJ or SJW} \\ \mu_{\text{drug}(i)} + \mu_{\text{study}(i)}, & \text{if no interaction} \end{cases}. \quad (13)$$

The logarithm of the magnitude of interaction was used to normalise its distribution. Simple and multiple linear regressions were applied (dependent variable being log GM ratio and independent variables being F and Fm). Correlation analyses were performed using Pearson's correlation coefficient. Results were considered significant for $p < 0.05$.

Bayesian hierarchical model-based meta-analyses

Model development

A hierarchical model was developed to meta-analyse statistically significant PK interaction data. Separate models were developed for each parameter of interest [$C_{\max}$, AUC (or 1/CL)] and described: (1) the ratios of (geometric) means for GFJ/SJW co-exposure *vs* exposure to single compounds and, (2) interindividual variability (as described by CV) specific to single compound exposure, GFJ and SJW co-exposure. Such model makes use of available PK data in a more robust manner than a standard weighted average of reported data through modelling reported means and population variability under a single model. This can be of particular relevance in cases under which interindividual variability may differ between population groups (here single exposure and co-exposure), as shown previously when comparing healthy adults and subgroups of the population (children, neonates, elderly) (Dorne et al. 2002, 2003a, b, 2004a, b).

Due to the log-normal nature of the data, input values for the hierarchical model were GM and geometric variance (GV) (or converted GM and GV) and sample sizes as reported in the publications. Data were modelled using normal distributions, after conversion of means and variances onto the log scale.

A meta-regression model was used to account for the relationship between the bioavailability F and the quantitative proportion of the compound handled by CYP3A4 (expressed as a percentage of the dose, Fm) and the interaction ratios. The meta-regression model assumed that (on log scale), the mean value μ_i is a sum of the effect of substrate, study and (interaction parameter γ) $\times$ (F + Fm). This "interaction parameter" was derived for both SJW and GFJ.

Observations (indexed by i) of GM, GV and sample size (n) were grouped by baseline drug, study, and occurrence of interaction (denoted int). Individual means and variances were converted onto the log scale and denoted LGM and LGV. Modelled mean values of LGM for observation i, μ_i, were meta-regressed as:

It follows that for a given drug (identified with its F and Fm values), the ratio of geometric means between the PK parameter for the baseline (single drug) and the PK parameter for the combined drug-GFJ or drug-SJW is of form $\exp[\gamma 1_{\text{int}(i)} F_i + \gamma 2_{\text{int}(i)} \text{Fm}_i]$, with $\gamma 1$ and $\gamma 2$ as the parameters to estimate (specific to GFJ and SJW). Assumption of linear relationship onto the log scale was formulated based on exploration of data (see "Relationship between substrate-specific parameters and magnitude of interaction").

Given the mean and sample size n, for each i, the observed values LGM and LGV are distributed as:

$$\text{GM}_i \sim \text{Normal}\left(\mu_i, \frac{\sigma_{\text{int}(i)}}{\sqrt{n_i}}\right), \quad (14)$$

$$\text{GV}_i \sim \text{Gamma}\left(\frac{n_i - 1}{2}, \frac{n_i}{2\sigma^2_{int(i)}}\right). \quad (15)$$

The scale parameter is thus made specific to the interaction data (none, GFJ, SJW) and assuming log-normal distributions, the coefficient of variation is also interaction-specific and given as:

$$\text{CV}_{\text{int}} = \sqrt{\exp(\sigma^2_{\text{int}} - 1)}. \quad (16)$$

In summary, the main model assumptions were (1) functional relationship between $C_{\max}$, AUC, 1/CL in mixture *vs* single compound, which depends on F and Fm, and (2), dependence of CV parameter only on whether interaction occurs.

The Bayesian model was coded and estimated with the Markov Chain Monte Carlo software Stan (Carpenter et al. 2017) using the Hamiltonian Monte Carlo approach. Convergence was checked after 1000 iterations. Vague and non-informative priors were used (Supplementary material S3).

Derivation of probabilistic uncertainty factors

CYP3A4-related UFs to cover the 95th and 99th percentiles of a population of healthy adults were calculated for each kinetic parameter and each study for GFJ and SJW, combining interindividual variability and magnitude of interaction, as described by Dorne et al. (2001a, b) for magnitude of change between subgroups of population.

Software

Statistical analyses were performed using R (version 3.3) and Stan (version 2.17). Data manipulation and graphical display were performed with R (tidyverse, broom and ggplot2 packages).

Results

Overview of data collected

Out of 1809 studies retrieved, relevant data from 109 publications were extracted providing 404 geometric mean ratios as magnitudes of interaction with GFJ and SJW for 89 different compounds. Specifically, for GFJ, 73 publications were included for 63 different compounds and a total of 257 mean ratios. For SJW, a total of 147 mean ratios were calculated from 36 publications dealing with 38 compounds. Statistically significant interaction data represented 62% of GFJ data and 64% of SJW data. The summary of the PK data collected on GFJ- and SJW-mediated interactions, together with information on data distribution, is presented in Fig. 1. The full dataset is provided in Supplementary material S2.

Figure 1 shows that the logarithms of GM ratios were mostly distributed around 0 according to inhibition or induction for GFJ and SJW, respectively (i.e., increased exposure with GFJ, decreased exposure with SJW). Few data were below 0 for GFJ and above 0 for SJW, illustrating an opposite effect (i.e., decreased exposure with GFJ, increased exposure with SJW). The resulting distributions describing the magnitudes of interactions were wider for GFJ compared with SJW, highlighting higher GM ratio values and variability. The overall variability was similar for acute and chronic PK parameters, for the baseline as well as the binary mixture data. In the same manner, the magnitudes of interactions following concomitant administration of GFJ or SJW were comparable for the acute and chronic PK parameters. Hence, data on acute and chronic parameters were pooled to assess the relationship between substrate-specific parameters and magnitude of interaction.

Relationship between substrate-specific parameters and magnitude of interaction

Oral bioavailability for individual substrates

The impact of substrate oral bioavailability (F) on the magnitude of interaction in the presence of CYP3A4 and/or Pgp inhibitor (GFJ) or inducer (SJW) was tested and the relationships are described in Fig. 2.

Figure 2 illustrates an inverse exponential relationship between F and magnitude of interaction for GFJ: the lower the substrate bioavailability, the higher the ratio between PK parameters measured with and without GFJ. A statistically significant ($p < 0.005$) medium correlation ($r = -0.44$) was described between F and the magnitude of interaction following GFJ co-exposure, for both the whole dataset and the statistically significant data. No statistically significant correlation was found between F and magnitude of interaction following co-exposure with SJW ($p = 0.06$ and 0.15 for the whole dataset and the statistically significant data, respectively).

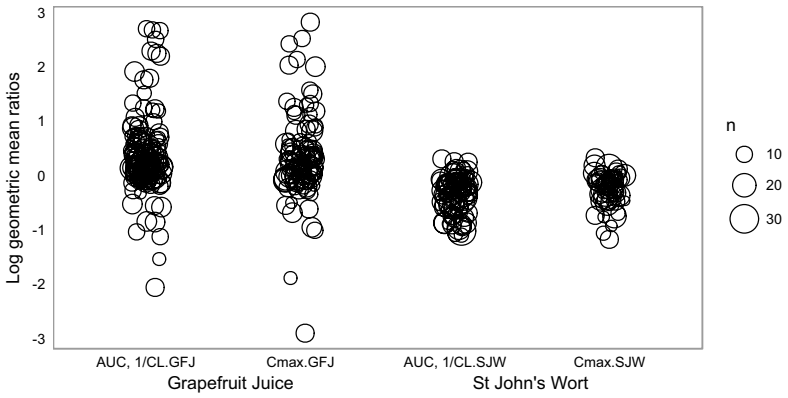


Fig. 1 Magnitude of inhibition or induction for markers of acute and chronic exposure after oral CYP3A4 and/or Pgp inhibition (GFJ) or induction (SJW) expressed as inter-study and inter-substrate variability in healthy adults. The magnitude of interaction is represented by the geometric mean ratio (log scale) between the mixture and baseline for a given parameter, for the same substrate and study. Each circle represents a dataset. The number of subjects n in the study is indicated by the size of the circle

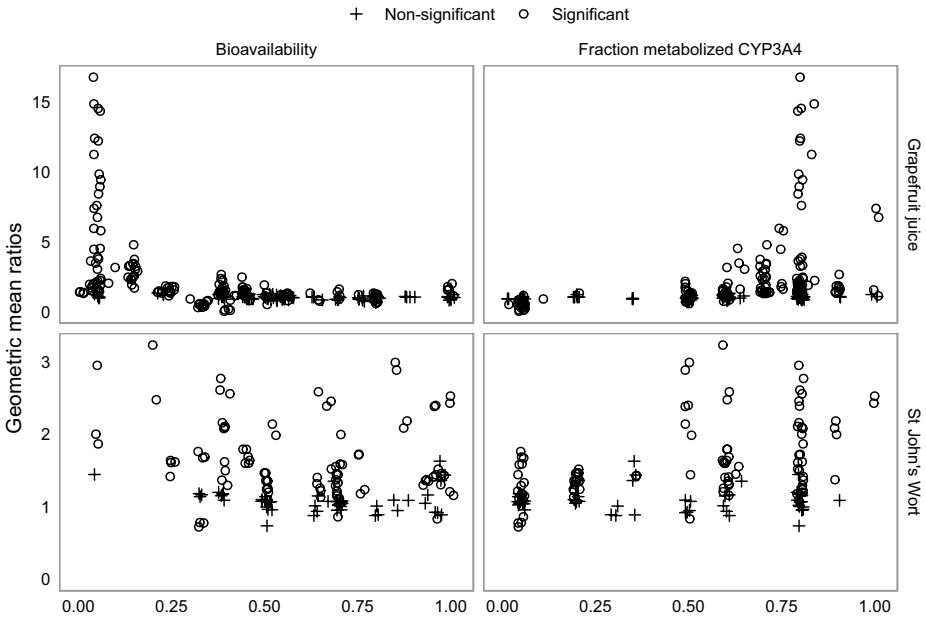


Fig. 2 Correlation between magnitude of interaction (for induction after SJW co-exposure, 1/ratio is represented), bioavailability (*x*-axis, left panel) and substrate fraction metabolised by CYP3A4 (*x*-axis, right panel) for both statistically (circles) and non-statistically (crosses) significant data

Quantitative fraction of substrate metabolised by CYP3A4

Substrates concomitantly exposed to GFJ were mainly CYP3A4 and/or Pgp substrates. Most substrates concomitantly exposed to SJW were CYP450 substrates, and in particular CYP3A4 substrates. The impact of the quantitative fraction metabolised by CYP3A4 (Fm) on the magnitude of interaction was investigated.

Figure 2 illustrates an exponential relationship between the extent of CYP3A4 metabolism and the magnitude of GFJ- and SJW-mediated interactions. This relationship highlights that the higher the CYP3A4 Fm for the substrate, the higher the magnitude of interaction ratio for PK parameters. A statistically significant ($p < 0.005$) medium correlation was described between CYP3A4 Fm and magnitude of interaction following GFJ co-exposure, for the whole dataset ($r = 0.32$) and the statistically significant data ($r = 0.36$). A statistically significant ($p < 0.005$) medium correlation was described between CYP3A4 Fm and the magnitude of interaction following SJW co-exposure, for the whole dataset ($r = 0.36$) and the statistically significant data ($r = 0.47$).

The impact of both explicative variables (i.e., substrate *F* and Fm) on the log values of the GM ratios was confirmed through a multiple linear regression analysis for all studies ($p < 0.005$ for both variables). No interaction between explicative variables was observed, suggesting independency of their impact on the magnitude of interactions. Results from these analyses on the relationship between substrate characteristics and magnitudes of interaction were integrated in the subsequent hierarchical model (see "Model development").

Predictive model using Bayesian hierarchical meta-analysis

Development of the meta-regression model

Parameter prior and posterior distributions after Bayesian calibration of the meta-regression model are described in Supplementary material S3. The posterior distributions of interaction parameters $\gamma 1$ and $\gamma 2$ well illustrate the impact of substrate characteristics *F* and Fm on magnitudes of interactions, in line with the relationships described in "Oral bioavailability for individual substrates" and "Quantitative fraction of substrate metabolised by CYP3A4".

The hierarchical Bayesian meta-regression model adequately predicted magnitudes of PK interactions as GM ratios between PK parameters for single compounds and

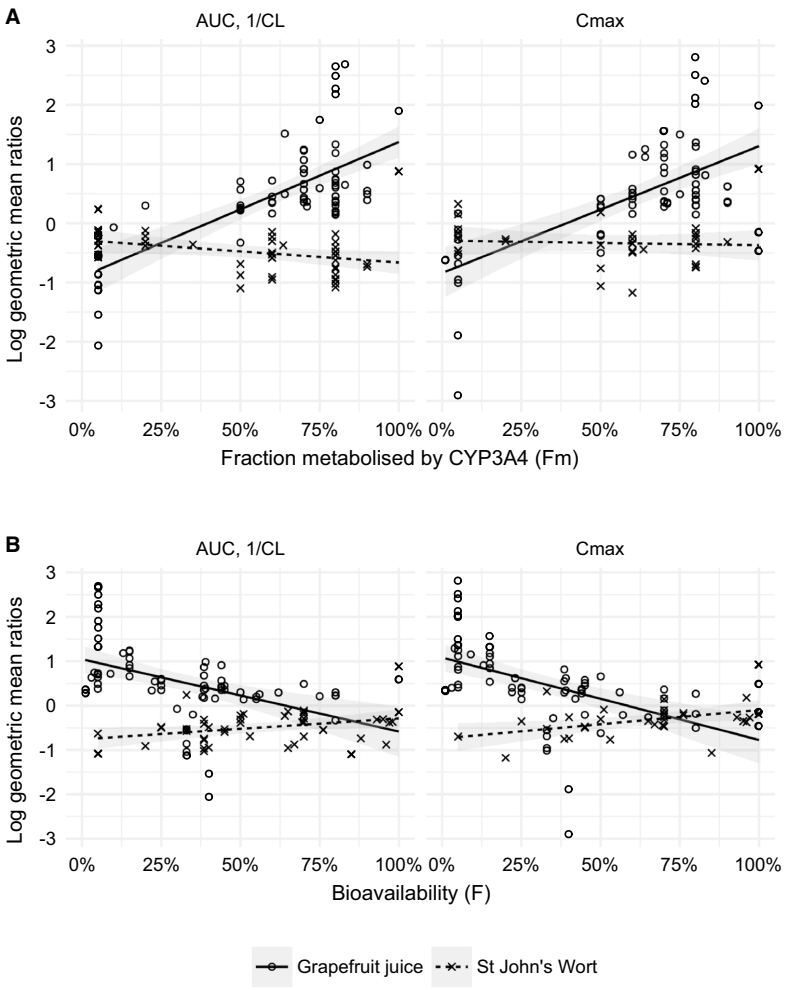


Fig. 3 Meta-regression model for CYP3A4 substrates predicting magnitudes of interaction (acute and chronic) in relation to fraction metabolised by CYP3A4 and bioavailability. Shaded line represents the most predictive linear trend in the dataset

PK parameters after co-exposure to GFJ or SJW, according to *F* and Fm. Summary of input data and model predictions is shown in Fig. 3.

Magnitudes of interactions

The meta-analysis with the Bayesian hierarchical model allowed the estimation of interaction parameters according to GFJ or SJW. For each model assumption, the magnitude of interaction following co-exposure with GFJ or SJW was estimated according to *F* and CYP3A4 Fm, as illustrated by Fig. 4 and exemplified in Supplementary material S4.

Results from the model predictions presented as three dimensional (Fig. 4) and in S4 reflect the relationships previously described between magnitudes of interactions after co-exposure with GFJ or SJW, *F* and CYP3A4 Fm. The highest magnitudes of interaction were characterised by ratios of [mean (95 CI)] 5.3 (4.7–5.9) for CYP3A4/Pgp inhibition by GFJ and 2.3 (2.1–2.6) for CYP3A4/Pgp induction by SJW.

Variability estimates

Interindividual variability was estimated using the hierarchical model for each PK parameter and each co-exposure scenario (GFJ and SJW), as reported in Table 1.

Overall interindividual variability for the CYP3A4 pathway was 52%. Interindividual variability following interaction was 58% and 42% for GFJ and SJW, respectively.

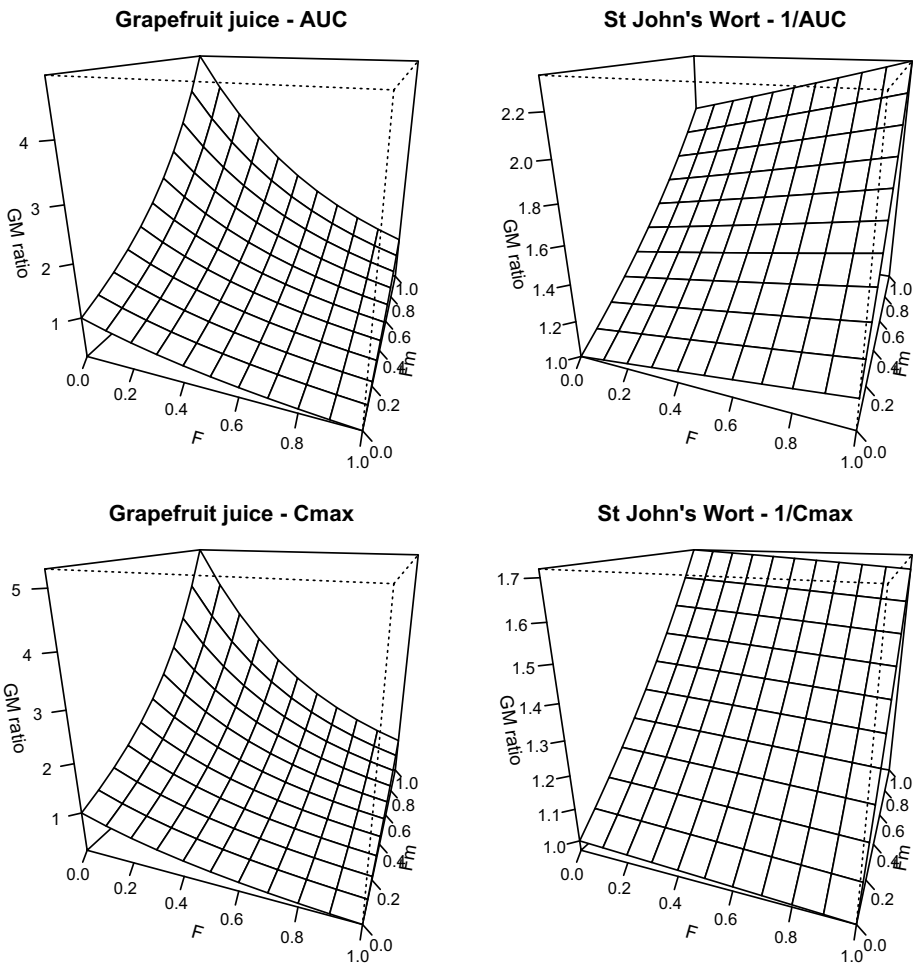


Fig. 4 Magnitudes of interactions (GM ratios) after CYP3A4 and P-glycoprotein inhibition and induction as a function of bioavailability (*F*) and substrate fraction metabolised by CYP3A4 (Fm)

Derivation of CYP3A4-related uncertainty factors for inhibition and induction

CYP3A4/Pgp pathway-related UFs for baseline (single compounds), inhibition (GFJ) or induction (SJW) were derived using the predicted magnitudes of interaction for a wide range of substrates from the meta-regression model as well as the associated variability for each set of parameters (C_{max}, AUC/CL). Results and distributions are provided in Supplementary material S5.

For single compounds, CYP3A4/Pgp variability was within the range of the default TK UF of 3.16 with CYP3A4/Pgp-related UFs between 2.3 and 3.4 (acute exposure) and 2.2–3.0 (chronic exposure) to cover the 95th and 99th percentiles of the population, respectively. When considering the maximum CYP3A4/Pgp inhibition, i.e., the highest fraction metabolised and the lowest bioavailability for GFJ data, CYP3A4/Pgp-related UFs values were up to sixfold higher than the default TK UF, ranging from 13.0 to 18.9 (acute exposure) and from 11.9 to 17.1 (chronic exposure) to cover 95th–99th percentiles of the human population. For CYP3A4/Pgp maximum induction by SJW, CYP3A4/Pgp-related UFs were similar to the default UF for acute exposure (range 3.1–4.0) and up to twofold higher for chronic exposure (range 4.9–6.7) to cover 95th–99th percentiles of the human population.

Table 1 Interindividual variability in markers of acute and chronic exposure for CYP3A4 substrates after co-exposure with Grapefruit Juice or St. John's Wort

N_p	N_d	N_{sub}	n	Interindividual variability CV_{pop} [mean (95 CI) %]
C_{max}				
72	Baseline 103	57	1192	55.6 (52.2–59.3)
51	GFJ 69	42	767	59.2 (54.3–64.4)
21	SJW 34	24	425	36.8 (33.8–40.1)
AUC or 1/CL				
86	Baseline 116	64	1350	50.5 (47.8–53.3)
56	GFJ 70	40	768	58.0 (53.6–62.5)
30	SJW 46	30	582	47.1 (43.7–50.7)

N_p = number of publications, N_d = number of datasets, n = number of subjects, N_{sub} = number of substrates, CV_{pop} = coefficient of variation (interindividual variability)

Discussion

Here, a meta-analysis on 109 human PK studies aimed to quantify oral CYP3A4 and Pgp inhibition (GFJ) and induction (SJW) for markers of acute and chronic exposure. Using the whole dataset and substrate characteristics (F, Fm), a Bayesian meta-regression model, integrating CYP3A4/Pgp PK variability, provided a range of quantitative estimates for CYP3A4 and Pgp induction and inhibition as the basis to derive CYP3A4/Pgp UFs for mixture risk assessment.

Mechanistic basis: key highlights

Overall, magnitudes of PK interactions for CYP3A4/Pgp substrates after inhibition (GFJ) or induction (SJW) were consistent with the known interaction mechanisms with CYP3A4/Pgp, namely competitive inhibition with GFJ leading to decreased elimination, and induction via pregnane X receptor (PXR) with SJW increasing overall elimination (Rahimi and Abdollahi 2012; Seden et al. 2010). However, in a few cases (see Supplementary material S2), inverse PK changes were observed for GFJ, which can be explained through inhibition of organic anion-transporting polypeptide (OATP)-mediated influx leading to an increase in elimination for some substrates (An et al. 2015; Bailey 2010; Hanley et al. 2011; Seden et al. 2010; Yu et al. 2017). It has to be noted that even if CYP3A4 substrates and inhibitors/inducers are known to overlap between with those of Pgp (Zhou 2008), inhibition of Pgp by GFJ in vivo has not been clearly demonstrated (Hanley et al. 2011). Additionally, GFJ inhibition of Pgp is short-lived (no more than 4 h) compared to the irreversible inhibition of CYP3A4 (Hanley et al. 2011; Seden et al. 2010), suggesting a longer and potentially higher inhibitory effect through CYP3A4 for most compounds. In the same way, even if SJW has also been described to induce other transporters and CYPs, the most robust evidence was available for CYP3A4 (Rahimi and Abdollahi 2012). Therefore, the current study can be considered to mainly deal with the CYP3A4 pathway.

The well-known impact of substrate characteristics on CYP3A4 inhibition by GFJ was further highlighted with a trend towards an increase in magnitude for substrates with low oral bioavailability through competitive inhibition of CYP3A4 in the gut (particularly the duodenum) (Gertz et al. 2008; Ohnishi et al. 2006; Takahashi et al. 2015). In addition, major CYP3A4 substrates (high Fm) showed higher magnitude of inhibition (GFJ) or induction (SJW) compared with minor substrates, in line with the mechanism-based inhibition (Seden et al. 2010) or induction (Rahimi and Abdollahi 2012), respectively.

Finally, the integration of bioavailability and CYP3A4 Fm in the model allowed to predict magnitudes of PK interactions for a wide range of substrates, while minimising confounding factors. A similar approach was used to predict PK interactions with GFJ elsewhere (Bailey 2017; Takahashi et al. 2015). The model-based predictions for a set of substrates were compared with an in vitro–in vivo extrapolation (IVIVE) approach proposed by Ainslie et al. (2014) for 15 substrates (Fig. 5). Predictions were within 25% of the observed magnitudes of interaction for 11 of the compounds, overestimated the magnitudes for three compounds (atorvastatin by 2.6-fold, methadone by 2.8-fold and sildenafil by 1.7-fold) and underestimated the magnitude for one compound (tacrolimus by 3.9-fold), illustrating an accurate and conservative approach. However, as the approach used in this work implies the knowledge of substrate (in vivo) characteristics, other tools such as in vitro methods/IVIVE or Quantitative Structure Activity Relationship (QSAR) models may be needed to predict interactions for unknown compounds (EFSA 2014). Furthermore, the main PK determinant in this study is CYP3A4 (either intestinal or hepatic) and this approach may be of limited use for compounds whose disposition is co-dependent on efflux/uptake transporters and metabolic enzymes. In the same manner, information on compounds plasma protein binding and systemic clearance relative to liver blood flow may be considered to discuss model predictions.

Interindividual variability

The hierarchical model provided refined estimates for CYP3A4-related interindividual variability in PK with an overall coefficient of variation (CV) of 52%, in line with

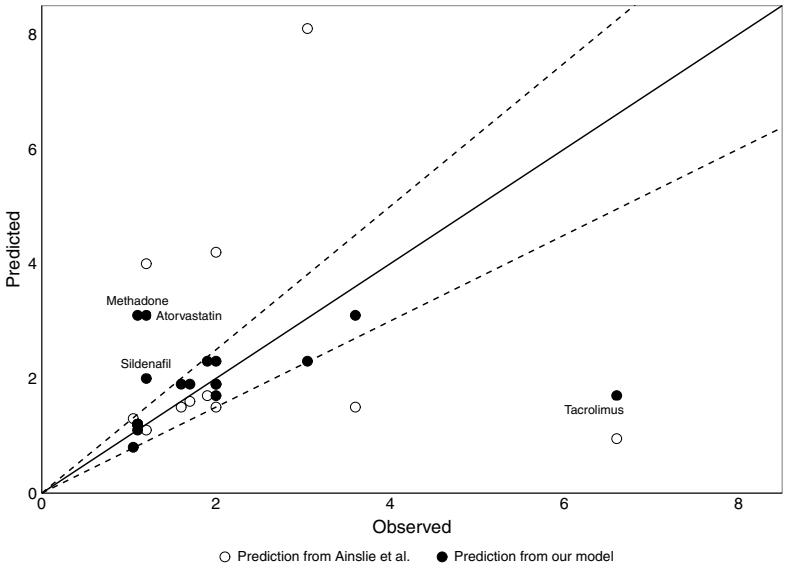


Fig. 5 Relationship between the predicted and observed AUC_{GFJ}/AUC ratios for 15 CYP3A4 substrates. The solid line denotes unity. Dashed lines denote 25% variability around the line of unity. Closed circles denote predictions from our model. Open circles denote predictions from Ainslie et al. (2014)

previous studies for which a CV of 46% was derived for oral exposure in healthy adults (Dorne et al. 2003a). Here, the slightly higher CV value can be explained by the wider range of substrates considered including minor CYP3A4 substrates and CYP3A4/Pgp substrates. In this case, inter-individual variability regarding expression and function of both CYP3A4 (Ince et al. 2013; Zhou 2008) and Pgp (Hoffmeyer et al. 2000; Kimchi-Sarfaty et al. 2007; Staud et al. 2010) should be accounted for.

Interindividual variability in PK parameters following CYP3A4/Pgp inhibition by GFJ was higher than that for single compounds with 58%. Extensive interindividual variability has been described for the expression of *CYP3A4* and *MDR1* genes (Lindell et al. 2003). Variability can also be due to the amount of GFJ ingested or the composition of GFJ in terms of flavonoids and furanocoumarins the main compounds involved as strong inhibitors in the PK interactions, with an inhibition constant K_i ranging 1–5 µM for 6′,7′-dihydroxybergamottin (Ainslie et al. 2014; Diaconu et al. 2011; Kawaguchi-Suzuki et al. 2017; Messer et al. 2012; Paine et al. 2005; Seden et al. 2010; Veronese et al. 2003).

Interindividual variability associated with SJW-mediated PK interaction was 42%. The variability in SJW-mediated CYP3A4 interaction has been described to be the consequence of polymorphisms on the *PXR* gene (Wang et al. 2009). In addition, variations can arise from various concentrations of hyperforin, one of the SJW constituents, inducing the transcription of *CYP3A4* and *ABCB1* genes (Mueller et al. 2006). The lower interindividual variability associated with SJW compared to GFJ could be explained, at least in part, by their main biological targets. Indeed, GFJ mainly targets intestinal CYP3A4, subject to extensive interindividual variability, whereas SJW targets both intestinal and hepatic CYP3A4 (Dresser et al. 2003).

CYP3A4-related uncertainty factors for inhibition and induction

Overall, the CYP3A4/Pgp UFs were within the default UF for TK (3.16) for single compounds (Dorne et al. 2003a). In contrast, for CYP3A4/Pgp inhibition or induction, UFs (95th–99th percentiles of the population) were up to sixfold and twofold higher compared to the default UF, respectively. Predicting PK interaction is critical in drug development since exposure to pharmaceuticals is often at the milligram (mg) level (de Boer et al. 2015), whereas in food safety, TK interactions in humans may be rare because exposure is often low (microgram level or less) and the inhibition/induction potential of the compound on the enzyme may not be reached (EFSA 2013). Hence, the range of CYP3A4/Pgp-related UFs for inhibition or induction that have been derived here from the pharmaceutical database, constitute conservative worst-case scenarios for CCI in a mixture risk assessment context. In practice, these CYP3A4/Pgp-related UFs could replace the default factor for human variability and the TK default factor for interspecies differences since no interspecies extrapolation would then be required for the TK dimension. Contexts of particular interest for mixtures in food safety may

include regulated chemicals and environmental contaminants particularly in the absence of human TK interaction data using in vitro evidence for CYP3A4/Pgp inhibition or induction as conservative in silico options. It is foreseen that as more in vitro data on TK interactions become available, this type of study can support the further development of quantitative IVIVE models for mixture risk assessment while integrating human in vitro and in vivo P/TK data on interactions with metabolic pathway-related variability.

Conclusion

Magnitude of GFJ- and SJW-mediated interactions, as well as human variability in kinetics for the CYP3A4 pathway, were estimated by a meta-regression model using a database of a wide range of compounds handled via this route. The predictive ability of the model in terms of magnitude of interaction was shown to be relatively accurate. However, this approach is more likely to be used in combination with other tools such as IVIVE or QSAR when information on the substrate P/TK is lacking, to increase the scope of applicability.

For a potential use of this approach in drug safety, it has to be reminded that only healthy adults were considered in this study. Therefore, the predicted magnitudes of interaction are likely to be different in specific subpopulations such as children and elderly (Ince et al. 2013; Klotz 2009), as well as for patients having various comedications (Xie et al. 2016) or impaired liver function (Almazroo et al. 2017).

Currently, there is a trend to replace traditional default or categorically based UFs using data-derived extrapolation factors, function of the population characteristics, dose metrics, exposure scenarios, metabolic and TK data and/or TD variability to reduce uncertainty in chemical risk assessment (Bhat et al. 2017). The model developed in this study and the results thereof illustrate the integration of human variability in metabolism and TK interactions in data-derived UFs. Results showed that CYP3A4-related UFs up to 19 and 6 (instead of the default kinetic factor of 3.16) would be required to cover up to 99% of individuals exposed to mixtures involving GFJ and SJW, respectively. Again, it has to be reminded that these constitute very conservative worst-case scenarios in a mixture risk assessment context.

Such approach, that could be applied to other metabolic pathways, can be used as a standalone tool for the (semi-quantitative) risk assessment process, or integrated in fully quantitative tools such as PB-P/TK models to inform higher tiers (Einolf 2007; EMA 2012; FDA 2012; Jamei 2016; Zhuang and Lu 2016).

Acknowledgements This work has been financed by the European Food Safety Authority (EFSA) under contract CFT/EFSA/EMRISK/2012/01 and Analytica LASER. The authors would like to thank Katarzyna Miernik, Iwona Kuter, Mateusz Nikodem, Agnieszka Zyla, Camille Béchaux, Sonia Halhol, Laure Perreau and Céline Dubuquoy from Analytica LASER for data collection and analysis.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Ainslie GR, Wolf KK, Li Y et al (2014) Assessment of a candidate marker constituent predictive of a dietary substance-drug interaction: case study with grapefruit juice and CYP3A4 drug substrates. J Pharmacol Exp Ther 351(3):576–584. https://doi.org/10.1124/jpet.114.216838

Almazroo OA, Miah MK, Venkataramanan R (2017) Drug metabolism in the liver. Clin Liver Dis 21(1):1–20. https://doi.org/10.1016/j.cld.2016.08.001

Ambudkar SV, Dey S, Hrycyna CA, Ramachandra M, Pastan I, Gottesman MM (1999) Biochemical, cellular, and pharmacological aspects of the multidrug transporter. Ann Rev Pharmacol Toxicol 39:361–398. https://doi.org/10.1146/annurev.pharmtox.39.1.361

An G, Mukker JK, Derendorf H, Frye RF (2015) Enzyme- and transporter-mediated beverage-drug interactions: an update on fruit juices and green tea. J Clin Pharmacol 55(12):1313–1331. https://doi.org/10.1002/jcph.563

Bailey DG (2010) Fruit juice inhibition of uptake transport: a new type of food–drug interaction. Br J Clin Pharmacol 70(5):645–655. https://doi.org/10.1111/j.1365-2125.2010.03722.x

Bailey DG (2017) Predicting clinical relevance of grapefruit–drug interactions: a complicated process. J Clin Pharm Ther 42(2):125–127. https://doi.org/10.1111/jcpt.12463

Bailey DG, Dresser GK (2004) Interactions between grapefruit juice and cardiovascular drugs. Am J Cardiovasc Drugs 4(5):281–297

Bhat VS, Meek MEB, Valcke M, English C, Boobis A, Brown R (2017) Evolution of chemical-specific adjustment factors (CSAF) based on recent international experience; increasing utility and facilitating regulatory acceptance. Crit Rev Toxicol 47(9):729–749. https://doi.org/10.1080/10408444.2017.1303818

Carpenter B, Gelman A, Hoffman MD et al (2017) Stan: a probabilistic programming language. J Stat Softw 76(1):1–32. https://doi.org/10.18637/jss.v076.i01

Choi JH, Ko CM (2017) Food and drug interactions. J Lifestyle Med 7(1):1–9. https://doi.org/10.15280/jlm.2017.7.1.1

de Boer A, van Hunsel F, Bast A (2015) Adverse food–drug interactions. Regul Toxicol Pharmacol 73(3):859–865. https://doi.org/10.1016/j.yrtph.2015.10.009

Diaconu CH, Cuciureanu M, Vlase L, Cuciureanu R (2011) Food–drug interactions: grapefruit juice. Rev Med Chir Soc Med Nat Iasi 115(1):245–250

Dorne JL, Walton K, Renwick AG (2001a) Uncertainty factors for chemical risk assessment: human variability in the pharmacokinetics of CYP1A2 probe substrates. Food Chem Toxicol 39(7):681–696

Dorne JL, Walton K, Renwick AG (2001b) Human variability in glucuronidation in relation to uncertainty factors for risk assessment. Food Chem Toxicol 39(12):1153–1173

Dorne JL, Walton K, Slob W, Renwick AG (2002) Human variability in polymorphic CYP2D6 metabolism: is the kinetic default uncertainty factor adequate? Food Chem Toxicol 40(11):1633–1656

Dorne JL, Walton K, Renwick AG (2003a) Human variability in CYP3A4 metabolism and CYP3A4-related uncertainty factors for risk assessment. Food Chem Toxicol 41(2):201–224

Dorne JL, Walton K, Renwick AG (2003b) Polymorphic CYP2C19 and N-acetylation: human variability in kinetics and pathway-related uncertainty factors. Food Chem Toxicol 41(2):225–245

Dorne JL, Walton K, Renwick AG (2004a) Human variability for metabolic pathways with limited data (CYP2A6, CYP2C9, CYP2E1, ADH, esterases, glycine and sulphate conjugation). Food Chem Toxicol 42(3):397–421. https://doi.org/10.1016/j.fct.2003.10.003

Dorne JL, Walton K, Renwick AG (2004b) Human variability in the renal elimination of foreign compounds and renal excretion-related uncertainty factors for risk assessment. Food Chem Toxicol 42(2):275–298

Dresser GK, Schwarz UI, Wilkinson GR, Kim RB (2003) Coordinate induction of both cytochrome P4503A and MDR1 by St. John's wort in healthy subjects. Clin Pharmacol Ther 73(1):41–50. https://doi.org/10.1067/mcp.2003.10

EFSA (2010) Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA J 8(6):1637. https://doi.org/10.2903/j.efsa.2010.1637

EFSA (2013) International framework dealing with human risk assessment of combined exposure to multiple chemicals. EFSA J 11(7):3313. https://doi.org/10.2903/j.efsa.2013.3313 **(69 pp)**

EFSA (2014) Modern methodologies and tools for human hazard assessment of chemicals. EFSA J 12(4):3638. https://doi.org/10.2903/j.efsa.2014.3638 **(87 pp)**

Einolf HJ (2007) Comparison of different approaches to predict metabolic drug–drug interactions. Xenobiotica 37(10–11):1257–1294. https://doi.org/10.1080/00498250701620700

EMA (2012) European Medicines Agency. Guideline on the investigation of drug interactions. Committee for Human Medicinal Products, London

FDA (2009) Food and Drug Administration. Guidance for industry: evidence-based review system for the scientific evaluation of health claims—final

FDA (2012) Food and Drug Administration. Guidance for industry: drug interactions studies: study design, data analysis, implications for dosing, and labeling recommendations. US Department of Health and Human Services, FDA, Silver Spring

Fujita K (2004) Food–drug interactions via human cytochrome P450 3A (CYP3A). Drug Metabol Drug Interact 20(4):195–217

Gertz M, Davis JD, Harrison A, Houston JB, Galetin A (2008) Grapefruit juice-drug interaction studies as a method to assess the extent of intestinal availability: utility and limitations. Curr Drug Metab 9(8):785–795

Hanley MJ, Cancalon P, Widmer WW, Greenblatt DJ (2011) The effect of grapefruit juice on drug disposition. Exp Opin Drug Metab Toxicol 7(3):267–286. https://doi.org/10.1517/17425255.2011.553189

Hennessy S, Leonard CE, Gagne JJ et al (2016) Pharmacoepidemiologic methods for studying the health effects of drug–drug interactions (DDIs). Clin Pharmacol Ther 99(1):92–100. https://doi.org/10.1002/cpt.277

Hoffmeyer S, Burk O, von Richter O et al (2000) Functional polymorphisms of the human multidrug-resistance gene: multiple sequence variations and correlation of one allele with P-glycoprotein expression and activity in vivo. Proc Natl Acad Sci USA 97(7):3473–3478. https://doi.org/10.1073/pnas.050585397

Ince I, Knibbe CA, Danhof M, de Wildt SN (2013) Developmental changes in the expression and function of cytochrome P450 3A isoforms: evidence from in vitro and in vivo investigations. Clin Pharmacokinet 52(5):333–345. https://doi.org/10.1007/s40262-013-0041-1

Isoherranen N, Kunze KL, Allen KE, Nelson WL, Thummel KE (2004) Role of itraconazole metabolites in CYP3A4 inhibition. Drug Metab Dispos 32(10):1121–1131. https://doi.org/10.1124/dmd.104.000315

Jamei M (2016) Recent advances in development and application of physiologically-based pharmacokinetic (PBPK) models: a transition from academic curiosity to regulatory acceptance. Curr Pharmacol Rep 2:161–169. https://doi.org/10.1007/s40495-016-0059-9

Kawaguchi-Suzuki M, Nasiri-Kenari N, Shuster J et al (2017) Effect of low-furanocoumarin hybrid grapefruit juice consumption on midazolam pharmacokinetics. J Clin Pharmacol 57(3):305–311. https://doi.org/10.1002/jcph.807

Kimchi-Sarfaty C, Oh JM, Kim IW et al (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315(5811):525–528. https://doi.org/10.1126/science.1135308

Klotz U (2009) Pharmacokinetics and drug metabolism in the elderly. Drug Metab Rev 41(2):67–76. https://doi.org/10.1080/03602530902722679

Kober M, Pohl K, Efferth T (2008) Molecular mechanisms underlying St. John's wort drug interactions. Curr Drug Metab 9(10):1027–1037

Lindell M, Karlsson MO, Lennernas H, Pahlman L, Lang MA (2003) Variable expression of CYP and Pgp genes in the human small intestine. Eur J Clin Invest 33(6):493–499

Messer A, Raquet N, Lohr C, Schrenk D (2012) Major furocoumarins in grapefruit juice II: phototoxicity, photogenotoxicity, and inhibitory potency vs. cytochrome P450 3A4 activity. Food Chem Toxicol 50(3–4):756–760. https://doi.org/10.1016/j.fct.2011.11.023

Mueller SC, Majcher-Peszynska J, Uehleke B et al (2006) The extent of induction of CYP3A by St. John's wort varies among products and is linked to hyperforin dose. Eur J Clin Pharmacol 62(1):29–36. https://doi.org/10.1007/s00228-005-0061-3

Naumann BD, Weideman PA, Dixit R, Grossman SJ, Shen CF, Sargent EV (1997) Use of toxicokinetic and toxicodynamic data to reduce uncertainties when setting occupational exposure limits for pharmaceuticals. Hum Ecol Risk Assess 3(4):555–565

Ohnishi A, Ohtani H, Sawada Y (2006) Major determinant factors of the extent of interaction between grapefruit juice and calcium channel antagonists. Br J Clin Pharmacol 62(2):196–199. https://doi.org/10.1111/j.1365-2125.2006.02636.x

Paine MF, Criss AB, Watkins PB (2005) Two major grapefruit juice components differ in time to onset of intestinal CYP3A4 inhibition. J Pharmacol Exp Ther 312(3):1151–1160. https://doi.org/10.1124/jpet.104.076836

Quignot N, Béchaux C, Amzal B (2015) Data collection on toxicokinetic and toxicodynamic interactions of chemical mixtures for human risk assessment. EFSA Support Publ 12(3):711E. https://doi.org/10.2903/sp.efsa.2015.EN-711

Rahimi R, Abdollahi M (2012) An update on the ability of St. John's wort to affect the metabolism of other drugs. Exp Opin Drug Metab Toxicol 8(6):691–708. https://doi.org/10.1517/17425255.2012.680886

Renwick AG, Lazarus NR (1998) Human variability and noncancer risk assessment- An analysis of the default uncertainty factor. Regul Toxicol Pharmacol 27(1 Pt 2):3–20. https://doi.org/10.1006/rtph.1997.1195

Roy K, Roy PP (2009) QSAR of cytochrome inhibitors. Exp Opin Drug Metab Toxicol 5(10):1245–1266. https://doi.org/10.1517/17425250903158940

Seden K, Dickinson L, Khoo S, Back D (2010) Grapefruit-drug interactions. Drugs 70(18):2373–2407. https://doi.org/10.2165/11585250-000000000-00000

Staud F, Ceckova M, Micuda S, Pavek P (2010) Expression and function of p-glycoprotein in normal tissues: effect on pharmacokinetics. Method Mol Biol (Clifton NJ) 596:199–222. https://doi.org/10.1007/978-1-60761-416-6_10

Takahashi M, Onozawa S, Ogawa R, Uesawa Y, Echizen H (2015) Predictive performance of three practical approaches for grapefruit juice-induced 2-fold or greater increases in AUC of concomitantly administered drugs. J Clin Pharm Ther 40(1):91–97. https://doi.org/10.1111/jcpt.12231

Veronese ML, Gillen LP, Burke JP et al (2003) Exposure-dependent inhibition of intestinal and hepatic CYP3A4 in vivo by grapefruit juice. J Clin Pharmacol 43(8):831–839

Wang XD, Li JL, Su QB et al (2009) Impact of the haplotypes of the human pregnane X receptor gene on the basal and St John's wort-induced activity of cytochrome P450 3A4 enzyme. Br J Clin Pharmacol 67(2):255–261. https://doi.org/10.1111/j.1365-2125.2008.03344.x

Won CS, Oberlies NH, Paine MF (2012) Mechanisms underlying food–drug interactions: inhibition of intestinal metabolism and transport. Pharmacol Ther 136(2):186–201. https://doi.org/10.1016/j.pharmthera.2012.08.001

Xie F, Ding X, Zhang QY (2016) An update on the role of intestinal cytochrome P450 enzymes in drug disposition. Acta Pharm Sinica B 6(5):374–383. https://doi.org/10.1016/j.apsb.2016.07.012

Yu J, Zhou Z, Tay-Sontheimer J, Levy RH, Ragueneau-Majlessi I (2017) Intestinal drug interactions mediated by OATPs: a systematic review of preclinical and clinical findings. J Pharm Sci 106(9):2312–2325. https://doi.org/10.1016/j.xphs.2017.04.004

Zhou SF (2008) Drugs behave as substrates, inhibitors and inducers of human cytochrome P450 3A4. Curr Drug Metab 9(4):310–322

Zhuang X, Lu C (2016) PBPK modeling and simulation in drug research and development. Acta Pharm Sinica B 6(5):430–440. https://doi.org/10.1016/j.apsb.2016.04.004

Chapter 4

An influenza model to inform vaccination policy in England

The work presented in this chapter is based on an influenza study conducted in two parts. First, an observational study was conducted to understand the current influenza burden and patterns of vaccination in the United Kingdom (using a retrospective analysis of observational data); second, we estimated the impact that vaccinating more children would have on the burden of influenza (using a Bayesian model). Output of the first part has been presented and published in Wiecek et al. (2015) and Rajaram et al. (2018), while the second part was described in Rajaram et al. (2017).[1] Here we are only concerned with presenting the second, modelling part of the study as this is where Bayesian modelling approaches were used.[2]

[1]The author's contribution was devising the model, implementation of all of the methods, statistical inference and co-writing the article.

[2]The published version of the article has been modified here to improve presentation and include additional information, which were originally part of supplementary material, in the article.

Impact of increased influenza vaccination in 2–3-year-old children on disease burden within the general population: a Bayesian model-based approach

Sankarasubramanian Rajaram[1], Witold Wiecek[2], Richard Lawson[3], Betina Blak[4], Yanli Zhao[5], Judith Hackett[3], Robert Brody[3], Vishal Patel[2], Billy Amzal[2]

[1]Formerly of AstraZeneca, Luton, UK; [2]LASER Analytica, London, UK; [3]AstraZeneca, Gaithersburg, MD, USA; [4]AstraZeneca, Luton, UK; [5]MedImmune, Gaithersburg MD, USA

Abstract

Introduction: During the 2013–2014 influenza season, Public Health England extended routine influenza vaccination to all 2- and 3-year-old children in England. To estimate the impact of this change in policy on influenza-related morbidity and mortality, we developed a disease transmission and surveillance model informed by real-world data.

Methods: We combined real-world and literature data sources to construct a model of influenza transmission and surveillance in England. Data were obtained for four influenza seasons, starting with the 2010–2011 season. Bayesian inference was used to estimate model parameters on a season-by-season basis to assess the impact of targeting 2- and 3-year-old children for influenza vaccination. This provided the basis for the construction of counterfactual scenarios comparing vaccination rates of ~2% and ~35% in the 2- and 3- year-old population to estimate reductions in general practitioner (GP) influenza-like-illness (ILI) consultations, respiratory hospitalizations and deaths in the overall population.

Results: Our model was able to replicate the main patterns of influenza across the four seasons as observed through laboratory surveillance data. Targeting 2- and 3-year-old children for influenza vaccination resulted in reductions in the general population of between 6.2–9.9% in influenza-attributable GP ILI consultations, 6.1–10.7% in influenza-attributable respiratory hospitalizations, and 5.7–9.4% in influenza-attributable deaths. The decrease in influenza-attributable ILI consultations represents a reduction of between 4.5% and 7.3% across all ILI consultations. The reduction in influenza-attributable respiratory

hospitalizations represents a reduction of between 1.2% and 2.3% across all respiratory hospitalizations. Reductions in influenza-attributable respiratory deaths represent a reduction of between 0.9% and 2.4% in overall respiratory deaths.

Conclusion: This study has provided evidence that extending routine influenza vaccination to all healthy children aged 2 and 3 years old leads to benefits in terms of reduced utilization of healthcare resources and fewer respiratory health outcomes and deaths.

1 Introduction

The World Health Organization estimates that influenza infection is responsible for between 3–5 million severe infections and 250,000–500,000 deaths globally each year [1]. In England and Wales, influenza infection is estimated to be responsible for between 7000–25,000 deaths during winter periods, with the highest mortality rates seen among persons aged 75 years and over [2]. It has been reported that children are predominantly responsible for the spread of influenza infection [3, 4], with a growing body of evidence suggesting that vaccinating healthy school children reduces the transmission of influenza [5, 6, 7] in the general population. In 2012, the Joint Committee on Vaccination and Immunisation (JCVI) issued a statement supporting the extension of routine influenza vaccination to all children aged 2–17 years in the United Kingdom (UK) [8]. The extension is being implemented by Public Health England (PHE) in a number of stages, the first stage of which was the vaccination of all 2- and 3-year-old children in the UK during the 2013–2014 influenza season. Additionally, during the 2013–2014 influenza season, geographical pilots in which influenza vaccination was offered to all 4- to 11-year-old children were implemented in seven distinct sites across England.

Estimating the seasonal burden of influenza is typically based on clinical surveillance systems which monitor respiratory healthcare outcomes and resource use including general practitioner (GP) visits, hospitalizations, and deaths. In the UK, PHE publish weekly surveillance reports that report rates of influenza-like-illness (ILI) GP consultations, respiratory hospitalizations, and all-cause mortality. In addition, since the H1N1 pandemic in 2009, the DataMart System has reported all major respiratory viral tests from a large number of laboratories across England. A number of studies have been published in the UK utilizing such data sources [9, 10, 11, 12, 2]. Time series methods have traditionally been used to estimate the influenza-attributable burden of non-specific outcomes such as GP consultations, hospitalizations, and deaths [9, 10]. One limitation associated with these approaches is the inability to estimate the population-level impact of vaccination; in particular, the impact of changing vaccination policies. Potential approaches to measure population-level impact include household trials of vaccinated and unvaccinated persons [13], geographical trials in which the entire population is randomized for vaccination [14, 15, 16], and surveillance methods to compare disease incidence prior to, and following the implementation of a vaccination policy [17]. However, each approach is associated with limitations of

external validity, and as such, disease transmission modeling approaches have been explored to estimate the impact of varying vaccination policies on influenza burden [18]. A notable study, connecting virologic data to a deterministic epidemiological model within Bayesian inference framework, was published by PHE in 2013 [19]. Models such as this one allow us to consider the impact of changing vaccination policy through an evaluation of observed data from recent seasons.

In this study we build on the approach described within the previous PHE study to estimate the impact of extending routine influenza vaccination to all 2- and 3-year-old children in England. Our approach is informed by a descriptive analysis of influenza-associated healthcare utilization and outcomes which has been previously published [20].

2 Methods

2.1 Model Population and demographics

The model was based on the total population in England as reported by the Office for National Statistics (ONS) at the mid-point of each of the four influenza seasons included in the study (2010–2011, 2011–2012, 2012–2013, and 2013–2014). A season was defined as beginning on 1 September and continuing until 13 April of the following year, which is in alignment with vaccination policy and the known notable period of influenza circulation [21, 22]. The model was informed by data on influenza vaccinations, ILI GP consultations, respiratory hospitalizations and deaths, and laboratory-confirmed virology surveillance data. All data inputs were age-specific according to the following age groups: 0–1 year olds, 2–3-year-olds, 4-year-olds, 5–10-year-olds, 11–17-year-olds, 18–64-year-olds and those aged 65 years and older. The under 18 age groups were selected to align with the anticipated age ranges for the roll-out of the UK childhood influenza immunization program, while the 65 and older age group were modeled separately as these subjects are routinely targeted for influenza vaccination. A short description of each data source and the relevant data inputs are provided within the sections below.

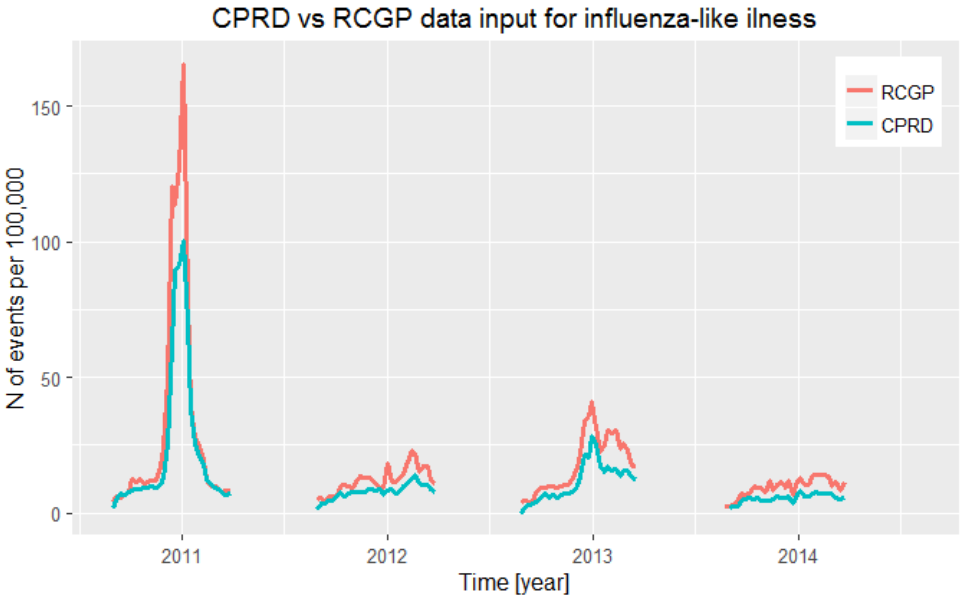
2.1.1 Vaccine Exposure Rate and Efficacy

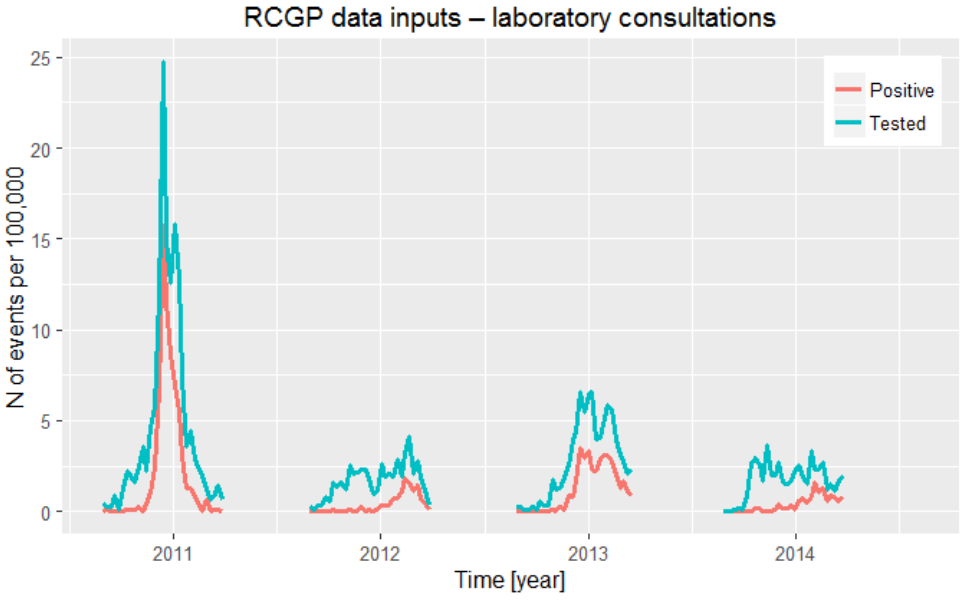
Age-specific vaccine exposure rates (for live attenuated influenza vaccine (LAIV) and trivalent influenza vaccine (TIV) were derived from a descriptive study undertaken using data from the Clinical Practice Research Datalink (CPRD) (Independent Scientific Advisory Committee (ISAC) protocol number 14_169RMn) for each season of the study. The rates observed within the CPRD have been demonstrated as representative of the vaccine exposure rates across England [23]. Vaccine efficacy has been shown to be 73% during seasons in which the vaccine is well matched to the dominant circulating influenza strain, and 46% during seasons in which there is a

mismatch [24]. It has also been shown that vaccine efficacy is 46% in elderly patients in comparison to 70% in younger patients [25]. Given that there was a good match between vaccine and circulating influenza strains observed during each of the seasons in the study, vaccine efficacy was assumed to be 70% for persons under 65 years, and 46% for persons aged 65 years and over, with equal efficacy assumed between LAIV and TIV [26].

2.1.2 ILI GP Consultations

ILI GP consultation data were obtained from the CPRD and the weekly returns service of the Royal College of General Practitioners (RCGP). The CPRD consists of routinely collected anonymized electronic healthcare record data from general practices across the UK. The patients in CPRD represent approximately 6.9% of the population of the UK and are considered to be broadly representative in terms of age, sex, and ethnicity of the population in England [27]. The weekly returns service of the RCGP monitors acute respiratory tract infections in England. The age and gender distribution of the RCGP surveillance network has been shown to be similar to that of the UK, with the only reported differences being a higher proportion of the population in the 25–44 age group and a lower proportion in the 0–4 age group [28]. Fig 1 illustrates weekly rates of ILI consultations derived from the CPRD and RCGP networks, which were included for each season in the model.


CPRD vs RCGP data input for influenza-like ilness
RCGP
CPRD
N of events per 100,000
150
100
50
0
2011
2012
2013
2014
Time [year]


RCGP data inputs – laboratory consultations
Positive
Tested
N of events per 100,000
25
20
15
10
5
0
2011
2012
2013
2014
Time [year]

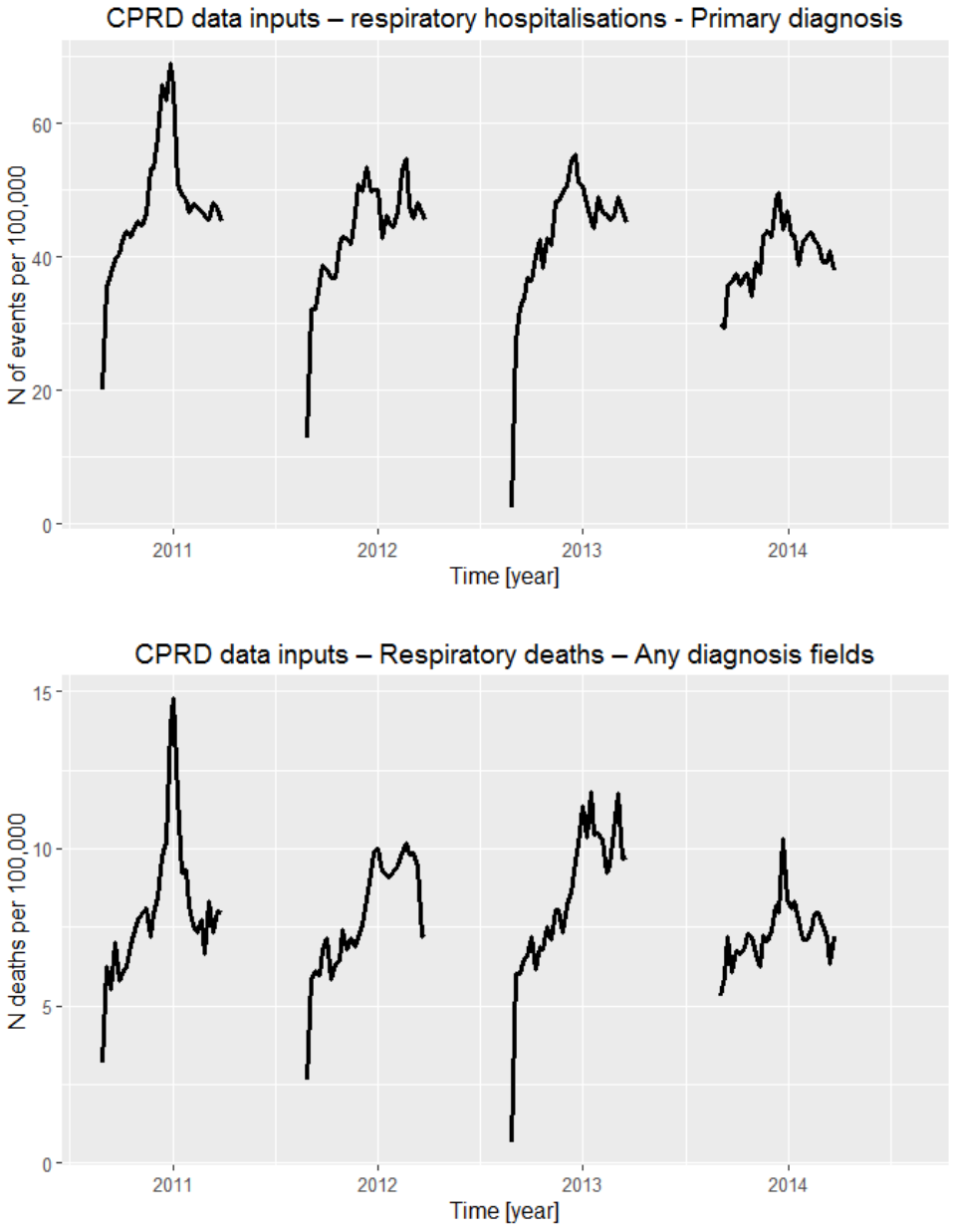


Fig 1. Influenza-like-illness consultation rates (CPRD and RCGP), all respiratory consultation rates (CPRD), respiratory hospitalization (HES), and respiratory deaths (ONS) data inputs for each season of the model. CRPD: Clinical Practice Research Datalink; RCGP: Royal College of General Practitioner; HES: Hospital Episode Statistics; ONS: Office for National Statistics.

2.1.3 Hospitalizations and Deaths

Rates of respiratory hospitalizations and respiratory deaths were obtained from the subset of patients within the CPRD who were eligible for linkage to the Hospital Episode Statistics (HES) and ONS databases, respectively (ISAC protocol number 14_169RMn). Respiratory hospitalizations were defined as any hospitalization with a respiratory ICD-10 code (J00-J99) or otitis media ICD-10 code (H65*, H66*, H67.1) listed as a primary diagnosis. Respiratory deaths were defined as any death record with a respiratory ICD-10 code (J00-J99) or otitis media ICD-10 code (H65*, H66*, H67.1) listed within any of the diagnosis fields. Fig 1 illustrates the weekly rates of respiratory hospitalizations and respiratory deaths for each season included in the study.

2.1.4 Laboratory-Confirmed Virology Surveillance Data

Virology data were obtained from the RCGP surveillance system and the Respiratory DataMart System (RDMS). Within the RCGP surveillance system, laboratory testing is undertaken for the majority of patients consulting for ILI. Influenza strain-specific (A/H1N1pdm, A/H1N1, A/H3N2, B/Victoria, and B/Yamagata) results were obtained for all persons who consulted for ILI and were tested for each season in the study.

The RDMS was established during the 2009 A/H1N1 pandemic as a laboratory-based surveillance system consisting of 14 PHE and National Health Service (NHS) laboratories in England [29]. Results from the RDMS surveillance system are reported within weekly PHE reports, published throughout the influenza season, and consist of the number of positive samples for a number of respiratory agents. The RDMS laboratory data was utilized to account for respiratory syncytial virus (RSV) infection within the model. Data on the total number of positive samples for RSV were extracted from the weekly reports for each of the seasons in the study.

2.1.5 Contact Information (POLYMOD)

To describe the age-specific mechanism of influenza spread, survey data from the POLYMOD study was used [30]. POLYMOD was a large scale study recruiting participants in eight countries, in which participants were asked to keep a diary of contacts accounting for age and type of contact. For the purpose of this study, physical contacts and those from UK were used, in accordance with the previously published model [19]. Based on 11,454 such contacts recorded in the study, a 7-dimensional square matrix M was created. Entry (i, j) in the matrix corresponds to the average daily number of contacts for an individual in age group i with individuals from age group j. Matrix M is illustrated in Fig 2.

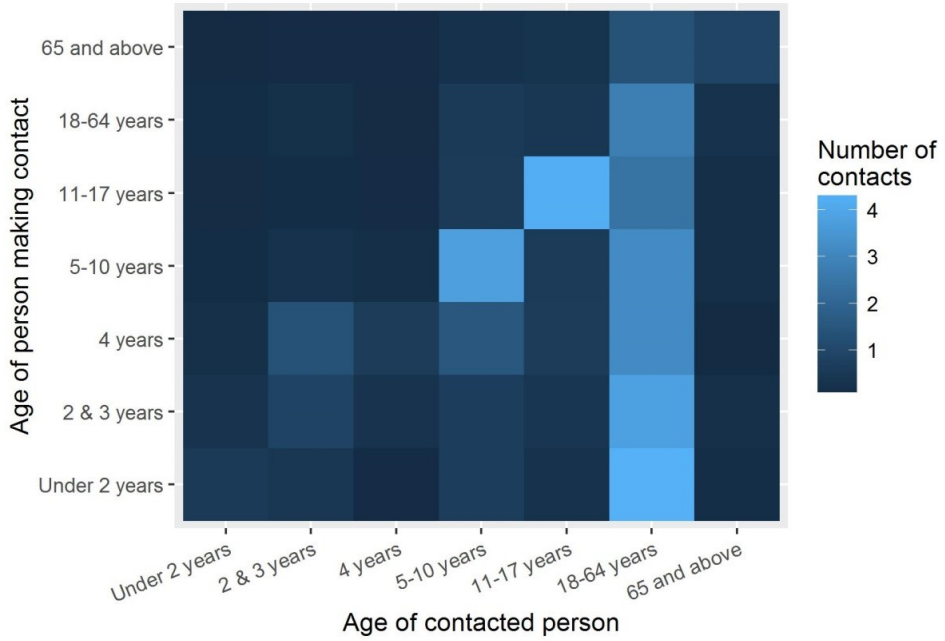


Fig 2. Matrix of age-specific contacts as derived from the POLYMOD study.

2.2 Model Overview

As outlined within the introduction, estimates of the reduction in influenza burden were based on varying vaccination rates, input into a deterministic disease model, with the resulting differences in burden compared. The model also allowed for estimation of sensitivity/specificity parameters, which link the modeled infection rates and healthcare utilization and outcomes (i.e., consultations, hospitalization, death, virological testing). Some of the model inputs (such as observed counts, vaccine effectiveness, vaccination rates, length of latent period) were fixed, while some were stochastic (parameters of SEIR model, sensitivity/specificity in the surveillance model of influenza).

Our goal was to use Bayesian inference to combine existing sources of data (as outlined above) and prior convictions (information from existing literature), to obtain posterior distributions for model parameters, that is, (i) stochastic inputs into the model and (ii) parameters linking the model to outcomes. Each influenza strain was modeled with a separate disease model, allowing for strain-specific virologic parameters, leading to different patterns of disease spread.

Using distributions of model parameters, we derived distributions of all quantities of interest, e.g., time series or seasonal total for ILI consultations or reproductive

numbers for the virus. Thanks to the use of the Bayesian approach it was possible to observe how uncertainty in model parameter estimates propagated into uncertainty in outcome estimates. Once the inference process was complete, vaccination rates (a fixed input into deterministic disease model) were manipulated to obtain estimates of burden for any hypothetical vaccination policy.

Bayesian inference was performed with the use of a Markov Chain Monte Carlo algorithm. Such process is computationally intensive, requiring use of a sufficiently fast algorithm for sampling from the posterior distribution.

To summarize, the inference model was partitioned into two main components: (i) a deterministic model generating underlying (latent) infection counts and (ii) an observational component linking them to modeled healthcare utilization and outcomes: consultations, hospitalizations, deaths, and virologic testing. The structure of the inference model is outlined in Fig 3. In the following sections we outline the structure of the deterministic disease model and surveillance model, followed by a description of model parameters and their prior distributions.

2.3 SEIR epidemiological model

The deterministic model used for inference was the same as described within the previous publication [19]: a Susceptible-Exposed-Infected-Resistant (SEIR) infectious disease model with an average latent period (γ_1) of 0.8 days and infectious period (γ_2) of 1.8 days. The population was divided into seven age groups as described earlier, with group sizes taken for the whole of England. As is standard in such SEIR models, we assumed random mixing within groups.

To account for pre-seasonal immunity in the population, susceptibility parameter σ_i was used, with i denoting an age group. Susceptibility modulated probability of infection for a member of group i by a season-specific, strain-specific constant. To avoid overfitting data, we assumed 4-parameter age groups for susceptibility: 0–4-year-olds, 5–17-year-olds, 18–64-year-olds, and those aged 65 years and over.

All subjects were assumed to be located in the susceptible compartment (S) at the beginning of the season, except for a small proportion (l) of those already infected. Sick individuals progressed through compartments E (exposed), I (infectious), and R (recovered/resistant). Two compartments were used for both E and I to obtain a more realistic gamma distribution of the duration of latent/infectious periods [31].

Vaccinated subjects were placed in the compartment R in a proportion reflecting the efficacy of the vaccine, with the rest remaining in their previous compartment. The probability of a vaccinated patient becoming immune did not change throughout the season. The flow of patients between compartments is illustrated in Fig 4.

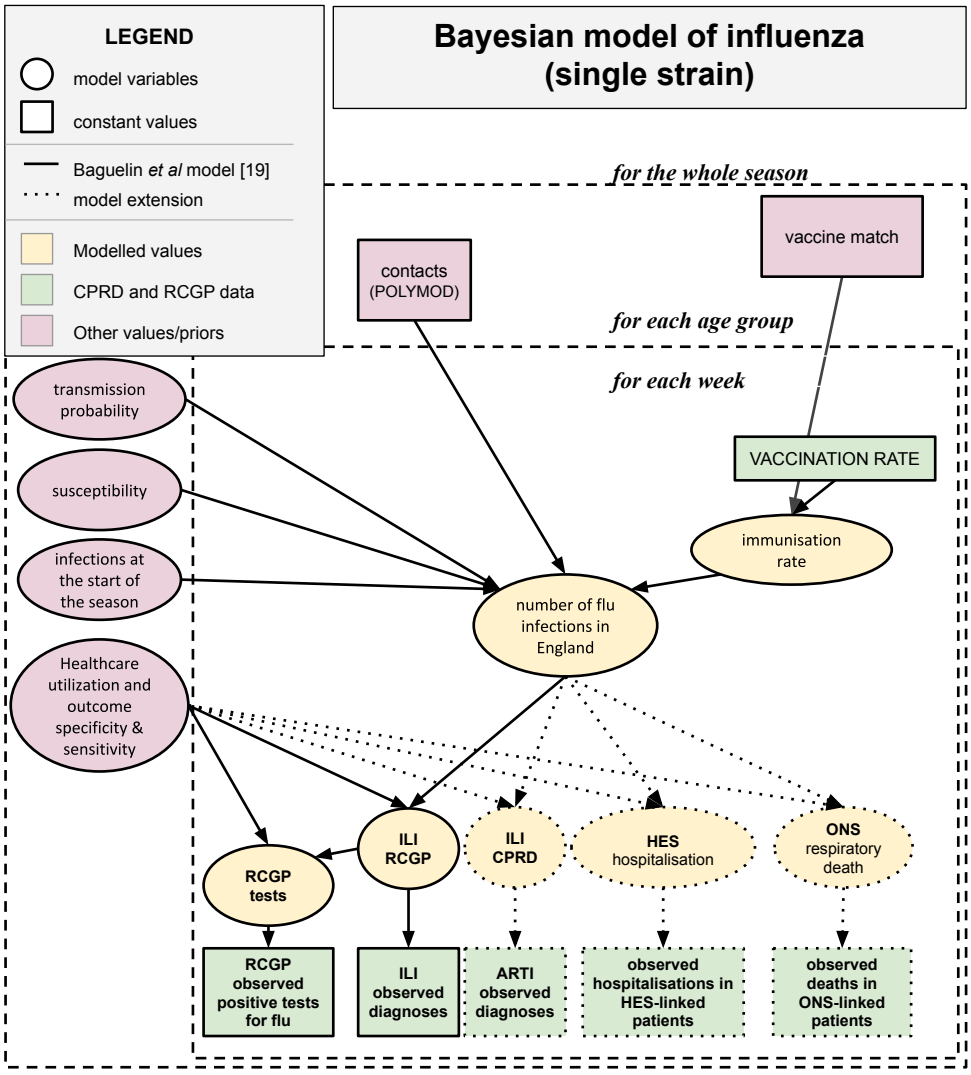


Fig 3. Structure of the inference model. ARTI: acute respiratory tract infections; CPRD: Clinical Practice Research Datalink; HES: Hospital Episode Statistics; ILI: influenza-like-illness; ONS: Office for National Statistics; RCGP: Royal College of General Practitioners.

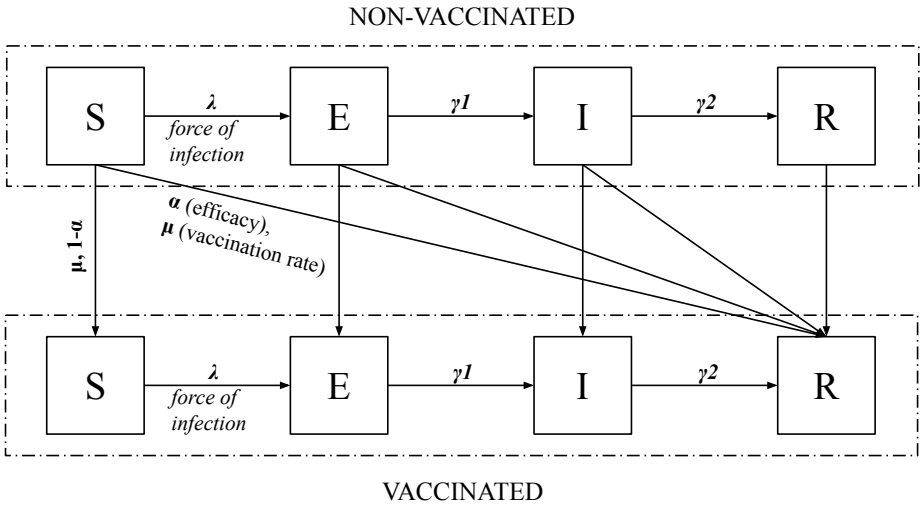


Fig 4. Schematic flow of patients between compartments (S-Susceptible, E-Exposed, I-Infecting, R-Resistant) for a single age group of the model. In model equations E and I compartments are further split into two compartments each.

Pre-existing resistance was described by the susceptibility parameter σ and varied between age groups. The probability of virus transmission when two susceptible individuals made contact (parameter q) was assumed to be constant for each strain within a given season.

Susceptibility (parameter σ), transmission probability (parameter q) and daily probability of contact with an infecting individual from a given age group (stored as matrix C), lead to deterministic force of infection for group i:

$$\lambda_i = q\sigma_i \sum_j \left(c_{\mathrm{ij}} I_j\right) ,$$

where j spans all age groups in the model and I_j is the number of infectious individuals in a given age group.

The state of the model could be determined at any given time by providing starting conditions (distribution of population between compartments) and using forward integration methods to update the system by a constant step. Number of infections during any week t for age group i was calculated as inflows into the "infected" compartments over appropriate number of steps. We denoted it by $z_i\left(t\right)$.

We noted that since force of infection was defined using multiplication of susceptibility and transmissibility parameters, the model was not identifiable if we assumed that all values of q and σ were equally likely. Therefore, the Bayesian model required an informative prior on the value of transmissibility to be identifiable.

Based on parameters of the SEIR model, the basic reproduction number R_0 was calculated as a dominant eigenvalue of the next-generation matrix [32], under the assumption of full susceptibility at the beginning of the season. Due to the role that age-specific susceptibility σ_i plays in force of infection λ_i, effective reproductive number R_e was considered, where susceptibility is modulated by values σ_i.

2.4 Surveillance Model

The surveillance model linked outputs of the deterministic model (counts of infections per age group and per week) with data on observed healthcare utilization and outcomes. For each utilization and outcome variable the aim was to assess its sensitivity and specificity to influenza. Out of $z_i(t)$ infected patients, only a fraction of patients would become symptomatic [33], and out of symptomatic patients only fractions would have a healthcare encounter (ILI consultation, hospitalization, death). These patients are recorded in CPRD and RCGP records. Furthermore, in RCGP some of the consulting patients can be tested for influenza. This situation is reflected in Fig 5. The goal was to relate z to various observed healthcare utilizations and outcomes via distributional assumptions.

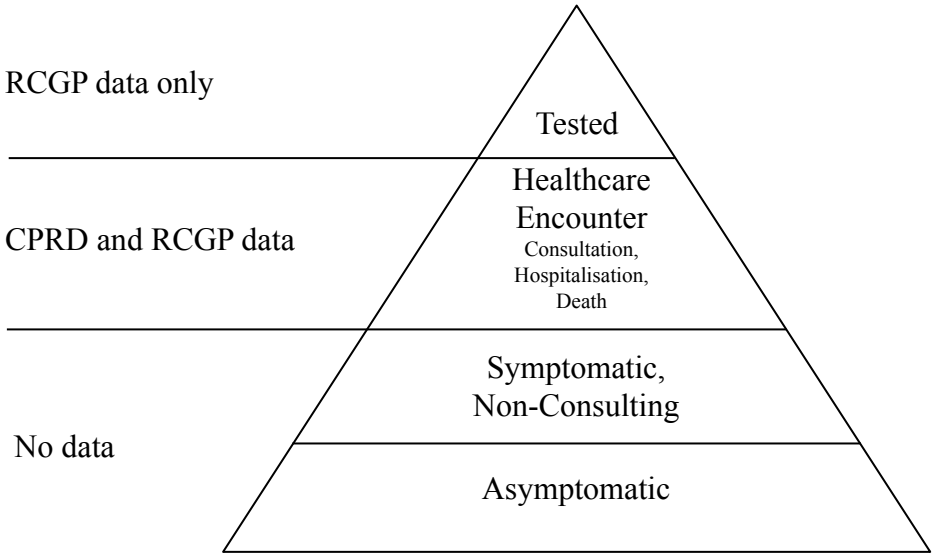


Fig 5. Surveillance pyramid of influenza: a schematic breakdown of influenza-infected population into groups. CPRD: Clinical Practice Research Datalink; RCGP: Royal College of General Practitioners.

For RCGP data on ILI consultations and subsequent laboratory confirmations of influenza, this link has already been described in the previous publication [19]. Availability of laboratory confirmations allowed us to employ a hypergeometric distribution making use of unobserved random variable m+ (number of influenza-positive patients)

and then marginalizing out the unknown parameter from posterior distribution. Constant sensitivity of each considered healthcare utilization and outcome throughout the season was assumed.

For CPRD healthcare utilization and outcomes – ILI GP consultations, hospitalizations and deaths – the lab confirmations were not routinely available; therefore, a simple distributional link was assumed, with counts for each age group I in week t assumed to follow a Poisson distribution. Denoting hospitalizations by h and deaths by d,

$$\begin{aligned} h_i(t) &\sim \text{Poisson}(\alpha_i^H z_i(t) + \beta_i^H RSV(t) + \gamma_i sin\left(\frac{2\pi}{52}(t-8)\right) + \delta_i^H) \\ d_i(t) &\sim \text{Poisson}(\alpha_i^D z_i(t) + \beta_i^D RSV(t) + \gamma_i sin\left(\frac{2\pi}{52}(t-8)\right) + \delta_i^D) \\ \text{ILI}_i(t) &\sim \text{Poisson}(\alpha_i^I z_i(t) + \delta_i^I) \end{aligned}$$

where α reflects $Pr(outcome|flu)$ (i.e., probabilities of hospitalization, death and, GP consultation for ILI), RSV(t) is a weekly number of lab-positive cases in DataMart that is (due to lack of data) not specific to age. Both in data exploration for this publication and in existing literature, lagging $RSV(t)$ by 2 weeks for children under 5 produced best fit to hospitalization data [34]. Therefore, for i corresponding to these groups, $RSV(t)$ was swapped for $RSV(t+2)$.

Parameter α can be understood as the proportion of infected patients who will present with a given outcome. β parameter does not have interpretation as a proportion but informs the within-season ratio of influenza/RSV-attributable outcomes, namely:

$$\text{(seasonal ratio of events due to flu vs due to RSV)} = \frac{\sum_t \alpha_i FLU_i(t)}{\sum_t \beta_i RSV(t)}$$

For α and ratio parameters informative priors were used, as outlined in the next section.

2.5 Priors

Susceptibility and transmissibility were the two crucial parameters for, one, death to hospitalization ratios for RSV and influenza, and two, sensitivity of healthcare utilization and outcomes. The sensitivity of CPRD ILI-related data was assumed to be proportional to sensitivity of ILI-related data in RCGP data, with log-normal prior on the proportion with low precision. Priors related to death and hospitalizations were derived from previous reports [9, 35] and are shown in Table 1. For priors relating to RCGP data the same priors as the PHE model publication [19] were used.

CPRD: Clinical Practice Research Datalink; ILI:– Influenza-like-illness; RCGP: Royal College of General Practitioners; RSV: respiratory syncytial virus.

Table 1. Prior values for model parameters.

Parameter	Age group						
	0–1	2–3	4	5–10	11–18	19–64	65+
ILI RCGP / ILI CPRD sensitivity	1						
Pr (hospitalization|influenza)	0.7%			0.002%		0.1%	8%
Pr (death|influenza)	0.001%						8%
Influenza vs RSV hospitalizations	0.15					1.8	1.2
Influenza vs RSV deaths	6						

2.6 Estimating impact of changing vaccination policy

Vaccine exposure rates were based on those observed within the CPRD population. To estimate the impact of targeting all 2- and 3-year- old children for routine vaccination on the morbidity and mortality burden of influenza in the general population, model estimates were derived and compared between two scenarios for each season.

3. **Scenario 1: Observed** vaccine exposure rates (for each age group for each season).
4. **Scenario 2: Modelled counterfactual scenario**: baseline vaccine exposure rates for each season + targeted vaccination of 2- and 3-year-old children for seasons 2010–2013, or without impact of targeted vaccination for the 2013–2014 season.

Comparing results between Scenario 1 and Scenario 2 for each season of the model enabled us to estimate the impact of targeted vaccination of 2- and 3-year-old children on influenza-associated morbidity and mortality (note, to estimate this impact for the 2013–2014 season, we reduced vaccine exposure rates from those observed in the 2- and 3- year-old population during that season to those that were observed during the 2012–2013 season, prior to the UK immunization program being implemented). Vaccine exposure rates used in the model for each scenario are provided within Table 2. It should be noted here that during the 2013–2014 season, an increase in vaccine exposure rates were also seen in children aged between 4 and 10 years. This is likely to be predominantly due to the misclassification of 3-year-old children within the 4-year-old age group.

When considering the scenario in which 2- and 3-year-old children were targeted for vaccination, we were interested in distributions for reductions in influenza-attributable ILI GP consultations, hospitalizations, and deaths compared to the values estimated by the inference model. To obtain rates of vaccination over time for Scenario 2, we assumed the same month-by-month distribution as is observed in Scenario 1, scaled by an appropriate constant. To create distributions for quantities of interest we used 5000 samples from the model posteriors and derived differences of interest via

the deterministic model with vaccination rates at two different levels (observed and hypothetical).

Table 2. Seasonal age-specific vaccine exposure rates for each scenario in the model, by season (obtained from Clinical Practice Research Datalink). Dots denote values unchanged from Scenario 1.

Season	Age group						
	0–1	**2–3**	**4**	**5–10**	**11–17**	**18–64**	**65+**
Scenario 1 (observed vaccine exposure rates)							
2010–2011	1.7%	3.1%	3.7%	3.9%	4.1%	10.2%	71.1%
2011–2012	0.7%	2.0%	2.7%	3.4%	4.0%	10.5%	71.1%
2012–2013	0.7%	1.8%	2.7%	3.4%	4.1%	10.7%	71.4%
2013–2014	0.7%	35.4%	14.9%	5.9%	4.3%	10.5%	70.7%
Scenario 2 (hypothetical vaccine exposure rates)							
2010–2011	.	35.4%	14.9%	5.9%	.	.	.
2011-2012	.	35.4%	14.9%	5.9%	.	.	.
2012-2013	.	35.4%	14.9%	5.9%	.	.	.
2013–2014	.	1.8%	2.7%	3.4%	.	.	.

2.7 Sensitivity Analysis

Available literature provides a wide range of estimates for balance between influenza- and RSV-attributable hospitalizations and deaths. Since the surveillance part of the model is highly dependent on its priors, a second analysis using an alternative set of values based on a recent publication was performed [34] (Table 3).

Table 3. Prior values for balance between influenza- and respiratory syncytial virus (RSV)-attributable hospitalizations and deaths used in sensitivity analysis.

	Age group (years old)			
	<5	**5–17**	**18–64**	**65+**
Hospitalizations due to influenza vs RSV	0.15	0.15	1.25	1.1
Deaths due to influenza vs RSV	1.1	1.1	1.1	1.4

To assess how strongly the model was determined by other priors, apart from considering different mean values for influenza/RSV ratio, precision on priors was decreased (by one third), to gauge propagation of uncertainty into burden and reduction estimates.

Another sensitivity analysis was performed by using both physical and non-physical contacts, but while the impact of including non-physical contacts did change estimates

of model parameters, the impact on the modeled outcomes was negligible and the results are not reported.

2.8 Statistical Inference

All data handling and preparation of outputs were done in R programming language, version 3.2.3. Core components of the model (computation of likelihood and evolution of deterministic SEIR model) were first coded in R and independently in C++ programming language to allow for quick computation.

The Markov Chain Monte Carlo (MCMC) algorithm used was an Adaptive Metropolis algorithm as described within the PHE model publication [19], extended to additional parameters present in our model (giving a total of 70 parameters).

Two chains of 3,000,000 iterations each were used, discarding the first million in a warm-up procedure. The algorithm was also tested on a simulated dataset, achieving convergence in less than 1,000,000 iterations. Computation for all four seasons was run in parallel and required approximately 24 hours of computational time on an Intel Core i5 2.3 GhZ processor.

3 Results

The model offered a good fit to considered healthcare utilization and outcomes of ILI, including GP consultations, hospitalizations, and deaths, within all seasons and age groups. Plots of fit of expected versus observed values for each group and season are presented in the as Figs 6–8. In certain cases (hospitalizations for 2011–2012, 2013–2014 seasons, deaths for 2013–2014 season), the peak of time series was not reproduced by the model. Such discrepancies can be a result of the peak being caused by factors not accounted for by the model or, more likely, disparities between individual data sources, namely virologic testing, GP consultations, hospitalizations, and deaths.

Dominant strains were A/H1N1 (pandemic) and B in 2010–2011, A/H3 in 2011–2012, B and A/H3 in 2012–2013, and A/H1N1 in 2013–2014. The 2010–2011 season was the most severe in terms of estimated number of infections.

3.1 Model Parameters

Estimated mean effective reproductive numbers for dominant strains were: 1.39 for A/H1N1 and 1.21 for B in 2010–2011, 1.16 for A/H3 in 2011–2012, 1.25 for B and 1.18 for A/H3 in 2012–2013, and 1.14 for A/H1N1 in 2013–2014. Pre-seasonal susceptibility estimates for dominating strains followed the same patterns as reported in the PHE model publication, with a pattern of low susceptibility in children and high susceptibility in adults [19].

For ILI GP consultations derived from CPRD we observed the same pattern in each season. The sensitivity of ILI in the infected population was higher in children under 5 and adults, than in children aged 5–17 years, and highest in the elderly, but with large season-to-season variability, e.g., for children aged under 5, mean values in four seasons were: 0.70%, 1.25%, 0.52%%, and 1.04%, respectively.

Similarly, for hospitalizations and deaths, variability was also observed. For the most at-risk cohort (over 65 population); the probability of death was 2.9%, 4.3%, 3.8%, and 8.7% in the four seasons, respectively, and the corresponding probability of hospitalization was 4.1%, 8.4%, 10.8%, and 9.7%, respectively. Such estimates are a result of balancing prior information (mean probabilities of hospitalization and deaths equal to 8%) and the model seeking best fit to observed data.

Posterior distributions of alpha and beta parameters of the model are presented in Supplementary Material Fig. S1–S6.[1]

3.2 Estimated Burden of Influenza

The model replicated the main patterns observed in laboratory data, namely high rates of B-type influenza in under 18-year-olds (seasons 2010–2011 and 2012–2013), A/H1 strains dominating in the 2010–2011 season, and B strains dominating in the 2012–2013 season. Table 4 details the strain-specific infection rate for each season. Temporal, age- and strain-specific patterns of infection are presented in Fig 9 (to improve readability the results have been summarized into four age categories).

Table 4. Infection rate by strain for each season in the study; dominating strain highlighted in bold.

Strain	Infection rate as a percentage of total population (95% credible interval)			
	2010–2011	**2011–2012**	**2012–2013**	**2013–2014**
A/H1N1pdm09	**30.2 (28.5–31.9)**	2.0 (1.6–2.4)	6.1 (4.6–7.5)	**4.5 (3.4–5.6)**
A/H1N1	3.6 (2.6–4.8)	0.0 (0.0–0.0)	7.3 (6.5–8.0)	0.5 (0.2–1.0)
A/H3	4.7 (3.8–5.6)	**5.0 (4.2–6.0)**	11.3 (9.8–12.7)	2.0 (1.4–2.7)
B	13.5 (12.2–15.0)	3.1 (2.3–4.0)	**20.7 (19.2–22.1)**	0.9 (0.4–1.4)

It was noted that the estimated infection rates were an order of magnitude lower in the elderly than in the rest of the population. This is a result of balancing of low counts in observed data, higher priors on sensitivity, and contacts derived from the contacts matrix in which the 65+ age group is the most isolated group.

Table 5 details rates of influenza-attributable ILI GP consultations, respiratory hospitalizations and respiratory deaths by season. Influenza-attributable ILI consultations

[1] We do not reproduce the additional plots and tables which are referenced in the publication as files S1 and S2. These are available, under DOIs: https://doi.org/10.1371/journal.pone.0186739.s001 for S1 and https://doi.org/10.1371/journal.pone.0186739.s002 for S2.

were highest during the 2010–2011 season (655.1 per 100,000 population) and lowest during the 2013–2014 season (76.8 per 100,000 population). The highest rates of influenza-attributable respiratory hospitalizations occurred during the 2012–2013 season (247.0 per 100,000 population) and the lowest rate occurring during the 2013–2014 season (31.1 per 100,000 population). The highest rates of influenza-attributable deaths occurred during the 2012–2013 season (80.0 per 100,000 population), and the lowest rates occurred during the 2011–2012 season (18.7 per 100,000 population).

Table 5. Influenza-attributable burden for influenza-like-illness (ILI) general practitioner (GP) consultations, respiratory hospitalizations, and respiratory deaths.

Outcome	**Mean flu-attributable cases per 100,000 population (95% credible interval)**			
	2010–2011	**2011–2012**	**2012–2013**	**2013–2014**
ILI GP consultations	655.1 (631.1–679.0)	168.9 (153.5–189.3)	316.3 (311.0–321.7)	76.8 (64.2–92.7)
Hospitalizations	114.4 (103.0–126.7)	46.3 (32.9–63.4)	247.0 (202.6–290.0)	31.1 (22.7–41.2)
Deaths	50.0 (43.6–56.7)	18.7 (10.2–28.7)	80.0 (56.9–104.7)	22.8 (13.8–32.0)

3.3 Estimated Reductions with Targeted Vaccination of all 2- and 3-year-old Children

Rates of influenza infection and rates of influenza-attributable ILI GP consultations, respiratory hospitalizations and deaths for the scenarios with and without targeted vaccination of 2- and 3-year old children are illustrated within Fig 10. It should be noted that targeted vaccination of 2- and 3-year-old children is represented as Scenario 2 for seasons 2010–2011, 2011–2012, and 2012–2013, and as Scenario 1 for 2013–2014 (the season during which this vaccination program was implemented by PHE).

Percentage reduction in age-specific infection rates are provided for each season within Table 6. The highest rates of reduction were seen in the 2- to 3-year-old group (reductions ranging between 26.8% to 29.0% across four seasons), followed by the 4-year-old age group (reductions ranging from 14.3% to 20.0% across the four seasons). Infection rates were reduced in the 0–1-year-old age group by between 7.2% (2010–2011 season) and 15.8% (2013–2014). Infection rates in age groups between 5–65+ years all saw reductions ranging from between 4.0% to 10.2% across the four seasons.

Reductions in influenza-attributable ILI GP consultations, respiratory hospitalizations and respiratory deaths are provided within Table 7. Across all seasons, influenza-attributable ILI consultations reduced by between 6.2% and 9.9%. Reductions in influenza-attributable respiratory hospitalizations were estimated to range between 6.1% and 10.7%, while reductions in influenza-attributable respiratory deaths ranged

between 5.7% and 9.4%. These reductions are provided in the context of all respiratory ILI consultations, respiratory hospitalizations, and respiratory deaths within Table 8. The decrease in influenza-attributable ILI consultations represents a reduction of between 4.5% and 7.3% across all ILI consultations. The reduction in influenza-attributable respiratory hospitalizations represents a reduction of between 1.2% and 2.3% across all respiratory hospitalizations. Reductions in influenza-attributable respiratory deaths represent a reduction of between 0.9% and 2.4% in overall respiratory deaths.

Table 6. Estimated percentage reduction in infection rate as a result of targeted vaccination of all 2- and 3-year-old children. The reduction values refer to total infections in Scenario 2 (Scenario 1 for 2013–2014) in comparison to Scenario 1 (Scenario 2 for 2013–2014) by age group due to higher vaccination rates in 2- and 3- year- old children.

Age group	**% reduction by season (95% CI)**			
	2010–2011	**2011–2012**	**2012–2013**	**2013–2014**
0–1	7.2 (6.4–8.3)	11.3 (7.5–16.2)	9.0 (5.8–13.6)	15.8 (9.8–25.2)
2–3	26.8 (26.1–27.7)	28.6 (25.3–33.0)	29.0 (26.4–32.7)	28.6 (23.4-36.8)
4	14.3 (13.6–15.3)	17.4 (13.8–22.0)	16.2 (13.3–20.3)	20.0 (14.6–28.4)
5–10	4.4 (3.6–5.3)	6.1 (3.9–9.2)	6.0 (4.0–8.3)	5.0 (2.9–7.9)
11–17	4.0 (4.8–3.2)	5.2 (3.3–7.9)	5.6 (3.9–7.8)	4.5 (2.4–7.0)
18–64	5.8 (4.7–7.1)	10.2 (6.5–14.7)	6.3 (4.2–8.9)	8.5 (5.6–13.1)
65+	5.7 (4.1–7.6)	9.3 (5.8–13.3)	6.4 (4.2–9.3)	9.4 (4.9–16.15)

Table 7. Reductions in influenza-attributable burden associated with targeted vaccination of 2- and 3-year-old children. Reductions represents the percentage by which influenza-attributable burden was reduced in Scenario 2 (Scenario 1 for 2013–2014) in comparison to Scenario 1 (Scenario 2 for 2013–2014) by age group due to higher vaccination rates in 2- and 3- year- old children.

Outcome	**% reduction in influenza-attributable cases (95% CI)** *Mean total reduction in number of influenza-attributable cases*			
	2010–2011	**2011–2012**	**2012–2013**	**2013–2014**
ILI GP	6.2 (5.0–7.4)	9.9 (6.5–14.0)	6.6 (4.6–9.2)	9.1 (6.1–13.7)
consultations	*21,084*	*8875*	*11,198*	*4153*
Hospitalizations	6.1 (4.6–7.6)	9.9 (6.4–14.1)	6.5 (4.3–9.4)	10.7 (6.2–17.8
	2223	*1500*	*5361*	*1251*
Deaths	5.7 (4.1–7.6)	9.3 (5.8–13.3)	6.4 (4.2–9.3)	9.4 (4.9–16.1)
	922	*569*	*1704*	*793*

Table 8. Percentage reduction in all respiratory outcomes associated with targeted vaccination of 2- and 3- year-old children. Reductions represent the percentage by which all-cause burden was reduced in Scenario 2 (Scenario 1 for 2013–2014) in comparison to Scenario 1 (Scenario 2 for 2013–2014) by age group due to higher vaccination rates in 2- and 3- year- old children.

Outcome	% reduction in respiratory outcomes (95% CI)			
	2010–2011	**2011–2012**	**2012–2013**	**2013–2014**
ILI GP cons.	5.7 (4.7–6.8)	7.3 (4.7–10.4)	6.3 (4.4–8.7)	4.5 (2.9–7.2)
Hospitalizations	1.2 (0.9–1.4)	1.2 (0.8–1.8)	2.3 (1.5–3.3)	1.3 (0.8–2.2)
Deaths	1.4 (1.0–1.9)	0.9 (0.4–1.6)	2.4 (1.3-3.7)	1.3 (0.6–2.4)

3.4 Sensitivity Analysis

As seen in parameter prior versus posterior distributions presented in the Supplementary Material, the ratios of influenza- to RSV-attributable cases was strongly dependent on assumed priors. This justifies the sensitivity analysis approach of exploring different prior values for these ratios.

Adjustment of priors for ratio parameters did not necessarily lead to lower estimates of morbidity and mortality burden, since in some cases estimated infection rates (or sensitivity parameters) could be adjusted upwards – we noted infection rates being adjusted higher in the 2013–2014 season, while other seasons estimates remained close to their "base case" values. Translated into the main outcome of respiratory healthcare utilization and outcome reductions, the sensitivity analysis resulted in the following ranges for influenza-attributable mean reductions over four seasons: 5.7–9.7% for ILI GP consultations, 5.3–10.1% in respiratory hospitalizations, and 4.8–10.1% in deaths – similar to the values presented in Table 7.

4 Discussion

As expected, the model estimates of morbidity and mortality burden in relation to infection rates indicated that the 2010–2011 season was associated with exceptionally high levels of influenza infection, attributable to the 2009 AH1N1 pandemic strain. The 2012–2013 season was also associated with high infection rates, with influenza B and AH3 strains dominating, with low activity also observed for the two H1N1 strains. The 2011–2012 and 2013–2014 seasons were both associated with low levels of infection. Overall trends observed across seasons correlated with those reported within PHE annual influenza reports [36, 37, 38, 39]. The model estimates for ILI GP consultations, respiratory hospitalizations, and respiratory deaths matched the weekly trends observed from CPRD and RCGP (ILI), HES (respiratory hospitalizations), and ONS (respiratory deaths).

4.1 Estimated Reduction in Burden Associated with Vaccination of 2- and 3-year-old Children

The estimated reductions in infections as a result of targeted vaccination of 2- and 3-year-old children were as expected across age groups, with the highest levels of reduction in the 2- to 3- and 4-year-old age groups (Table 6). These age groups are those that we would expect to benefit most from direct protection as they were targeted for vaccination. Reductions in infection rates were similar across the other age groups, although children under the age of 1 experienced greater reductions in infection rates. Evaluating across seasons demonstrated that reductions were larger for the less severe seasons (2011–2012 and 2013–2014). The reductions were also subject to a higher level of uncertainty during less severe seasons, as evidenced by generally wider confidence intervals when compared with severe seasons.

There were similar reductions in the rates of influenza-attributable ILI GP consultations, respiratory hospitalizations, and respiratory deaths within each season (Table 7). The trends across seasons followed a similar trend to infection rates, with greater reductions in the two less severe seasons in comparison to the more severe seasons for each of the outcomes. Table 8 demonstrates how the reductions in influenza-attributable healthcare utilization and outcomes translate into reductions in overall respiratory outcomes. There is less variation in reductions across seasons, due to cases attributable to other circulating pathogens diluting the effect of influenza vaccination (except for ILI consultations, which are highly sensitive to influenza). The impact on overall respiratory outcomes has been provided, as they may be easier to interpret than influenza-attributable outcomes, which are rarely reported at a population level in public health surveillance (virologic surveillance is routinely reported; however, laboratory-confirmed hospitalizations, deaths, and GP ILI consultations are not).

4.2 Comparing Model Estimates with External Estimates of Burden

Comparing our model estimates of morbidity and mortality burden with published estimates derived using traditional time series approaches is challenging, as few studies reporting on the seasonal burden of influenza in the UK have been published since the 2009 H1N1 pandemic. A 2016 publication by Matias et al. reported seasonal rates of influenza-attributable hospitalizations and deaths based on data collected from HES and ONS between 1997 and 2009 [12]. Our model estimates of influenza-attributable hospitalizations (ranging from 34.6–239.3 per 100,000) were higher than those reported within the Matias et al. publication (mean seasonal rate of 48 per 100,000 population). Likewise, for influenza-attributable deaths, our model estimates (ranging from 15.1 to 49.9 per 100,000) were higher than those reported by Matias et al. (mean seasonal rate of 12 per 100,000). However, it is challenging to determine whether the difference is caused by the difference in methodology or differences in surveillance data following the 2009 H1N1 pandemic (e.g., introduction of the respiratory DataMart system for virologic in 2009 [29]), which are used for both types

of models. In contrast to typical regression time series analysis, the model we have developed is highly determined by an assumed model of transmission. However, such an underlying model is necessary to be able to estimate reduction in burden for counterfactual scenarios – something that is not possible in a time series approach. PHE annual influenza publications report on laboratory-confirmed intensive care unit and high-dependency unit admissions through the UK severe influenza surveillance system. Our model-estimated trends of influenza-attributable hospitalizations across seasons compare well with those reported by PHE, except for the 2012–2013 season, where they are much higher than those reported by PHE in comparison to the other four seasons [39]. This may be reflective of the unusually long period of influenza circulation that occurred during 2012–2013, which was also characterized by influenza B circulating prior to influenza A.

In 2014, PHE published results estimating the impact of the regional pilot program implemented during the 2013–2014 season, in which school children between the ages of 4 and 11 years were targeted for vaccination [40]. PHE collected and compared data between pilot and non-pilot regions, including data on ILI GP consultations and laboratory-confirmed influenza-attributable hospitalizations. The results did not reach statistical significance; however, the vaccination program had an estimated impact of 66% on ILI consultations, and 24% on influenza-attributable hospitalizations [40]. In comparison, our model estimated an impact of vaccination of 4.5% on ILI consultations and 10.7% in influenza-attributable respiratory hospitalizations, although these estimates are based on the targeted vaccination of 2- and 3-year-old children only. While a direct comparison of the results is not possible, it is encouraging to see that the type of reductions estimated by our model could be observed through routinely collected surveillance data.

4.3 Sensitivity of Model Results

Individual model parameters were impacted by prior distributions (see Supplementary Material) as evidenced by sensitivity analysis. However, due to the combination of multiple data sources, the results were resilient to changes in parameters, and reductions observed in sensitivity analysis were close to reductions estimated in base case scenario.

Further exploration is needed into the sensitivity of results on model assumptions, especially surrounding the surveillance model and the way that sensitivity of outcomes is modeled. We acknowledge that further exploration will be needed to assess the impact of these assumptions on reduction estimates and propose a number of alternatives to explore in the section that follows.

4.4 Model Strengths and Limitations

In this study, we observed similar susceptibility patterns, effective reproductive values and sensitivity patterns of ILI as the previously published model [19]. However, this

similarity in itself does not validate the approach, as our approach also shares some of the limitations of that model. Most importantly, disconnect between modeled seasons did not allow us to model acquired immunity and requires estimation of age-specific susceptibility profiles solely from data insufficient for the task. Secondly, assumptions of constant values of parameters throughout the season (especially sensitivity) might not be a good approximation of how people behave during severe epidemics. Thirdly, vaccination is treated as having a simplified dichotomous effect of fully immunizing a vaccinated person or not at all, while in some cases the vaccination may solely modulate the patient's susceptibility to influenza.

This model combines the approach of modeling underlying influenza transmission together with a range of healthcare utilization and outcomes in England including GP consultations, deaths and hospitalizations. The models for data obtained from CPRD are parsimonious, with simple distributional assumptions, accounting only for the presence of one time-dependent (and non-age-specific) covariate (RSV). Such a simplified model of healthcare utilization and outcomes could in the future be refined by more complicated distributional assumptions. Combined with large number of parameters such a model can result in overfitting to data – in this case observed in fits to ILI GP consultation data. An ideal model, for example, would benefit from a linking between outcomes (hospitalizations and deaths) and virological testing, which allows use of hypergeometric distribution.

For the purposes of the model, vaccine efficacy was assumed to be the same between nasal and injectable vaccines, with good match for all considered seasons and strains. In reality, factors such as seasonal strain drift and differential effectiveness across strains could lead to higher variability in vaccine efficacy. The recent publication [41] around reduced efficacy of LAIV for H1N1 is not part of the analyses, which may be a limitation; however, the assumption of same efficacy between LAIV and IIV will negate any material difference in the results.

Parameters related to outcomes are also estimated on a season-by-season basis, which can lead to overfitting to data. We attempted to summarize this inter-seasonal variability and gauge the impact of prior information via a sensitivity analysis. Another approach could be to estimate parameters jointly while still allowing them to vary across seasons. Such approach could explicitly model acquired immunity, allowing for the treatment of susceptibility as a dynamic parameter.

The approach we described here can be extended to additional influenza-related outcomes, including broader and non-specific respiratory GP outcomes, as declining rates of ILI have been observed over the last 10 years [42]. However, this requires a model capable of accounting for other circulating viruses or comprehensive virological surveillance data.

5 Conclusion

The findings of our model support the claim that extending routine influenza vaccination to all healthy 2- and 3-year-old children leads to benefits in terms of reduced utilization of healthcare resources and fewer respiratory health outcomes and deaths within the general population

Competing Interests

SR is a former employee of AstraZeneca. RL, BB, JH, and RB are employees of AstraZeneca. YZ is an employee of MedImmune, the biologics arm of AstraZeneca. BB holds shares in AZ. WW, BA, and VP are employees of LASER Analytica, who have received funding for the current study from AstraZeneca. On behalf of all authors, these commercial affiliations do not alter our adherence to PLOS ONE policies on sharing data and materials.

Funding

This study was supported by AstraZeneca. The funder provided support in the form of salaries for authors BB, RB, JH, RL, YZ, and SR (during the time of the study when SR was employed by AstraZeneca) and funding to Laser Analytica to conduct the study. Members of AstraZeneca were permitted to review the manuscript and offer comments, but the authors decided whether or not to address these comments. AstraZeneca did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section

Acknowledgements

Modelling expertise and guidance was provided by Richard Pitman. Clinical guidance and expertise was provided by Douglas Fleming.

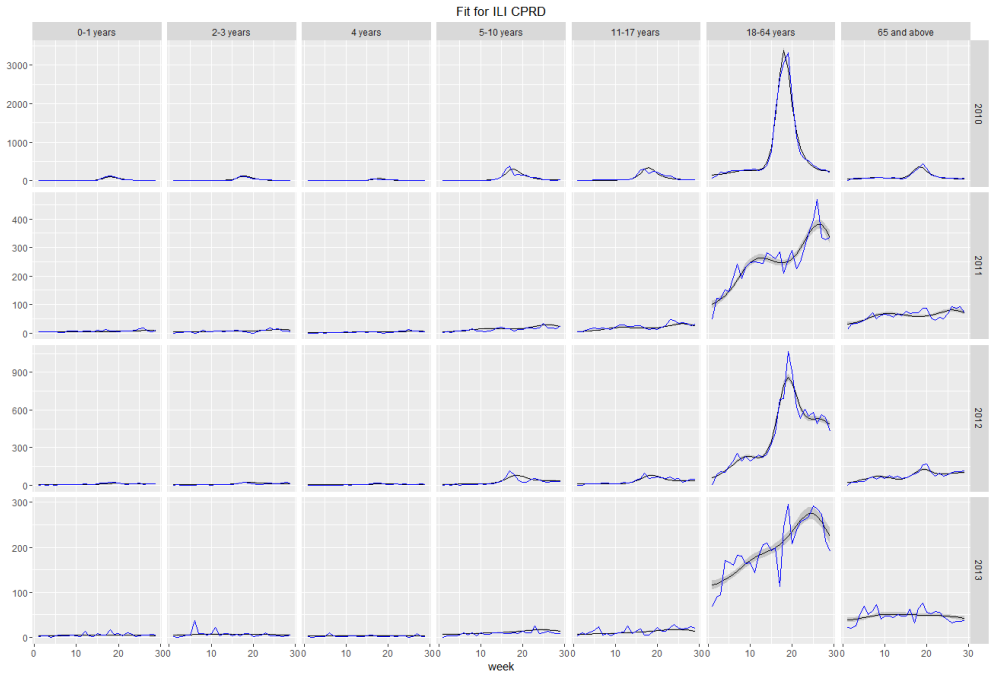


Fig 6. Model-estimated number of weekly influenza-like-illness (ILI) consultations (grey line) vs Clinical Practice Research Datalink observed number of weekly ILI consultations (blue line)

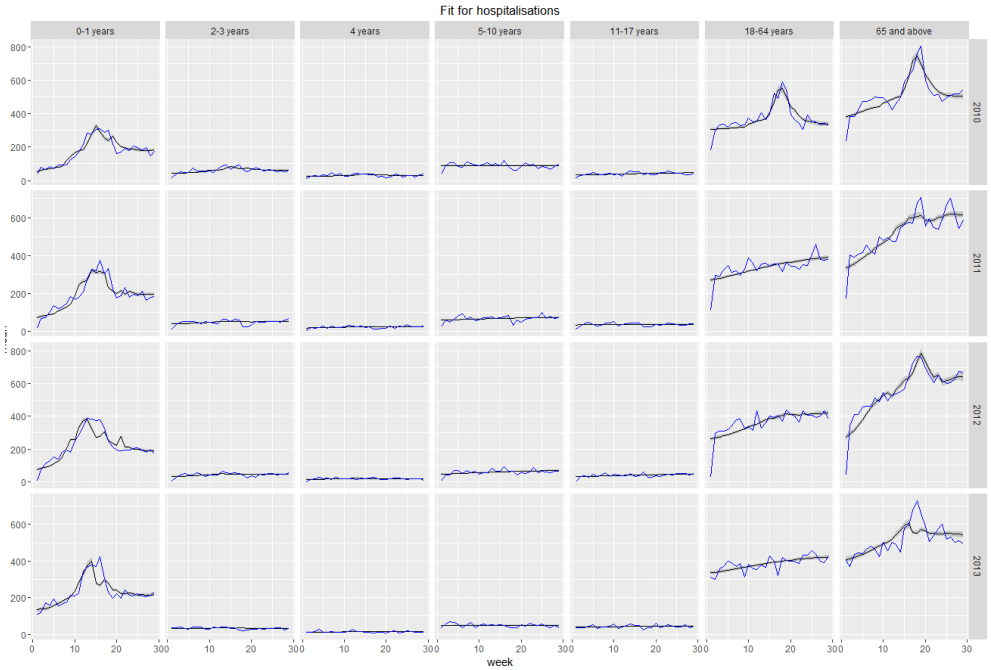


Fig 7. Model-estimated number of weekly respiratory hospitalizations (grey line) vs observed number of weekly respiratory hospitalizations (blue line)

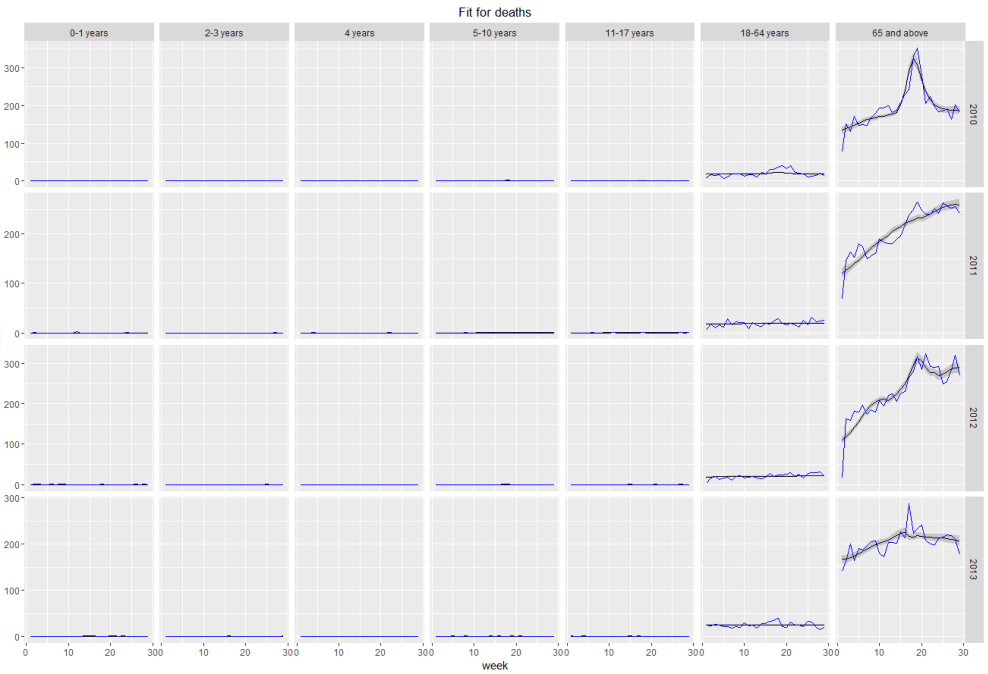


Fig 8. Model-estimated number of weekly respiratory deaths (grey line) vs observed number of weekly respiratory deaths (blue line)

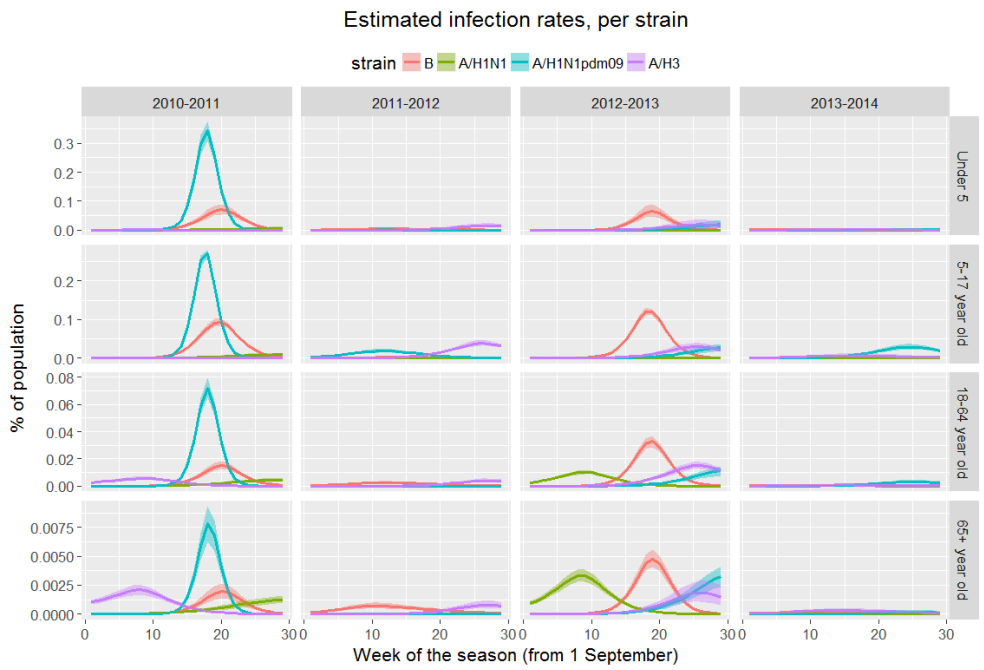


Fig 9. Strain-specific weekly infections for each season, by age group.

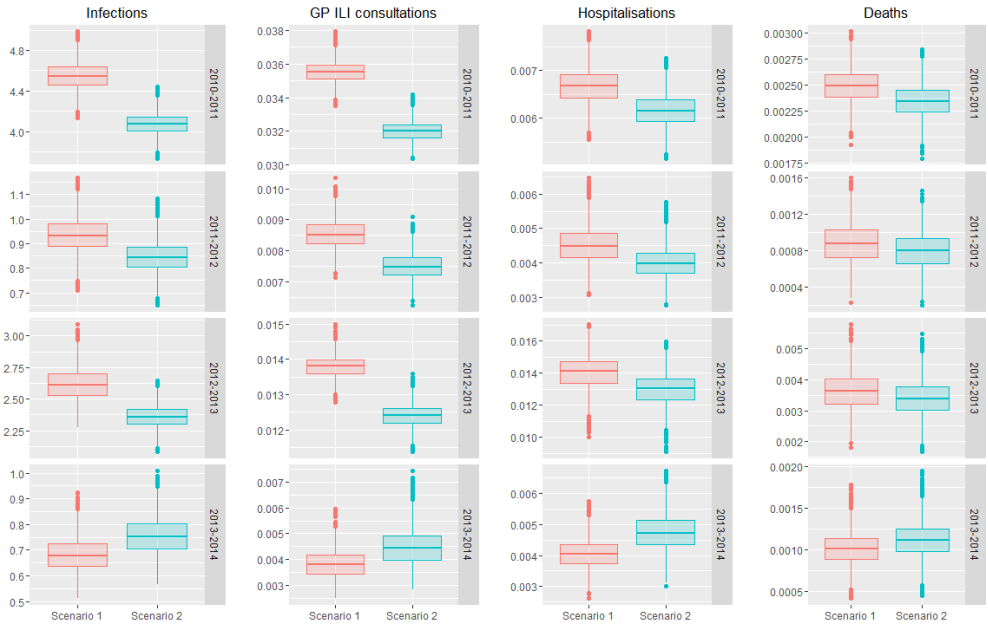


Fig 10. Comparison of rates of influenza infection and influenza-attributable burden between observed (Scenario 1) and modelled (Scenario 2) values. Horizontal bar is the median, with shaded bar ("hinges") representing 25th and 75th percentiles. Vertical bar spans values within 1.5 times inter-quantile range from hinges. In seasons 2010-2013 the modelled rates are with targeted vaccination. In 2013-2014 lack of targeted vaccination is the model.

References

[1] WHO. Influenza (Seasonal). https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal), 2014.

[2] P. Hardelid, R. Pebody, and N. Andrews. Mortality caused by influenza and respiratory syncytial virus by age group in England and Wales 1999-2010. *Influenza and Other Respiratory Viruses*, 7(1):35–45, January 2013.

[3] Simon Cauchemez, Alain-Jacques Valleron, Pierre-Yves Boëlle, Antoine Flahault, and Neil M. Ferguson. Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature*, 452(7188):750–754, April 2008.

[4] Cécile Viboud, Pierre-Yves Boëlle, Simon Cauchemez, Audrey Lavenu, Alain-Jacques Valleron, Antoine Flahault, and Fabrice Carrat. Risk factors of influenza transmission in households. *The British Journal of General Practice: The Journal of the Royal College of General Practitioners*, 54(506):684–689, September 2004.

[5] A. S. Monto, F. M. Davenport, J. A. Napier, and T. Francis. Effect of vaccination of a school-age population upon the course of an A2-Hong Kong influenza epidemic. *Bulletin of the World Health Organization*, 41(3):537–542, 1969.

[6] Derek Weycker, John Edelsberg, M. Elizabeth Halloran, Ira M. Longini, Azhar Nizam, Vincent Ciuryla, and Gerry Oster. Population-wide benefits of routine vaccination of children against influenza. *Vaccine*, 23(10):1284–1293, January 2005.

[7] T. A. Reichert, N. Sugaya, D. S. Fedson, W. P. Glezen, L. Simonsen, and M. Tashiro. The Japanese experience with vaccinating schoolchildren against influenza. *The New England Journal of Medicine*, 344(12):889–896, March 2001.

[8] The Joint Committee on Vaccination and Immunisation. JCVI statement on the annual influenza vaccination programme – extension of the programme to children 2012. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/224775/JCVI-statement-on-the-annual-influenza-vaccination-programme-25-July-2012.pdf, 2012.

[9] R. J. Pitman, A. Melegaro, D. Gelb, M. R. Siddiqui, N. J. Gay, and W. J. Edmunds. Assessing the burden of influenza and other respiratory infections in England and Wales. *The Journal of Infection*, 54(6):530–538, June 2007.

[10] D. M. Fleming. The contribution of influenza to combined acute respiratory infections, hospital admissions, and deaths in winter. *Communicable Disease and Public Health*, 3(1):32–38, March 2000.

[11] D. M. Fleming, R. J. Taylor, F. Haguinet, C. Schuck-Paim, J. Logie, D. J. Webb, R. L. Lustig, and G. Matias. Influenza-attributable burden in United Kingdom primary care. *Epidemiology and Infection*, 144(3):537–547, February 2016.

[12] Gonçalo Matias, Robert J. Taylor, François Haguinet, Cynthia Schuck-Paim, Roger L. Lustig, and Douglas M. Fleming. Modelling estimates of age-specific influenza-related hospitalisation and mortality in the United Kingdom. *BMC public health*, 16:481, August 2016.

[13] Susanna Esposito, Paola Marchisio, Samantha Bosis, Lara Lambertini, Laura Claut, Nadia Faelli, Ciro Bianchi, Giorgio L. Colombo, and Nicola Principi. Clinical and economic impact of influenza vaccination on healthy children aged 2-5 years. *Vaccine*, 24(5):629–635, January 2006.

[14] Pedro A. Piedra, Manjusha J. Gaglani, Claudia A. Kozinetz, Gayla Herschler, Mark Riggs, Melissa Griffith, Charles Fewlass, Matt Watts, Colin Hessel, Julie Cordova, and W. Paul Glezen. Herd immunity in adults against influenza-related illnesses with use of the trivalent-live attenuated influenza vaccine (CAIV-T) in children. *Vaccine*, 23(13):1540–1548, February 2005.

[15] Mark Loeb, Margaret L. Russell, Vanessa Manning, Kevin Fonseca, David J. D. Earn, Gregory Horsman, Khami Chokani, Mark Vooght, Lorne Babiuk, Lisa Schwartz, Binod Neupane, Pardeep Singh, Stephen D. Walter, and Eleanor Pullenayegum. Live Attenuated Versus Inactivated Influenza Vaccine in Hutterite Children: A Cluster Randomized Blinded Trial. *Annals of Internal Medicine*, 165(9):617–624, November 2016.

[16] Mark Loeb, Margaret L. Russell, Lorraine Moss, Kevin Fonseca, Julie Fox, David J. D. Earn, Fred Aoki, Gregory Horsman, Paul Van Caeseele, Khami Chokani, Mark Vooght, Lorne Babiuk, Richard Webby, and Stephen D. Walter. Effect of influenza vaccination of children on infection rates in Hutterite communities: A randomized trial. *JAMA*, 303(10):943–950, March 2010.

[17] Beate Sander, Jeffrey C. Kwong, Chris T. Bauch, Andreas Maetzel, Allison McGeer, Janet M. Raboud, and Murray Krahn. Economic appraisal of Ontario's Universal Influenza Immunization Program: A cost-utility analysis. *PLoS medicine*, 7(4):e1000256, April 2010.

[18] Mark Jit, Anthony T. Newall, and Philippe Beutels. Key issues for estimating the impact and cost-effectiveness of seasonal influenza vaccination strategies. *Human Vaccines & Immunotherapeutics*, 9(4):834–840, April 2013.

[19] Marc Baguelin, Stefan Flasche, Anton Camacho, Nikolaos Demiris, Elizabeth Miller, and W. John Edmunds. Assessing optimal target populations for influenza vaccination programmes: An evidence synthesis and modelling study. *PLoS medicine*, 10(10):e1001527, October 2013.

[20] W. Wiecek, B. Amzal, S. Bakshi, V. Patel, and T. Van Staa. Age Related Consultation Rates of Clinically-Diagnosed Influenza And Acute Respiratory Illnesses Observed Through A Network of Gp Practices Across England. *Value in Health*, 18(7):A579, November 2015.

[21] Department of Health. The flu immunisation programme 2013/14 – extension to children 2013, July 2013.

[22] Public Health England. PHE weekly influenza report 2014. https://www.gov.uk/government/statistics/weekly-national-flu-reports, 2014.

[23] Sankarasubramanian Rajaram, Amy Steffey, Betina Blak, Matthew Hickman, Hannah Christensen, and Herve Caspard. Uptake of childhood influenza vaccine from 2012-2013 to 2014-2015 in the UK and the implications for high-risk children: A retrospective observational cohort study. *BMJ open*, 6(8):e010625, January 2016.

[24] Tom Jefferson, Carlo Di Pietrantonj, Alessandro Rivetti, Ghada A. Bawazeer, Lubna A. Al-Ansary, and Eliana Ferroni. Vaccines for preventing influenza in healthy adults. *The Cochrane Database of Systematic Reviews*, (7):CD001269, July 2010.

[25] Tom Jefferson, Carlo Di Pietrantonj, Lubna A. Al-Ansary, Eliana Ferroni, Sarah Thorning, and Roger E. Thomas. Vaccines for preventing influenza in the elderly. *The Cochrane Database of Systematic Reviews*, (2):CD004876, February 2010.

[26] D. M. Fleming, N. J. Andrews, J. S. Ellis, A. Bermingham, P. Sebastianpillai, A. J. Elliot, E. Miller, and M. Zambon. Estimating influenza vaccine effectiveness using routinely collected laboratory data. *Journal of Epidemiology and Community Health*, 64(12):1062–1067, December 2010.

[27] Emily Herrett, Arlene M. Gallagher, Krishnan Bhaskaran, Harriet Forbes, Rohini Mathur, Tjeerd van Staa, and Liam Smeeth. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International Journal of Epidemiology*, 44(3):827–836, June 2015.

[28] Ana Correa, William Hinton, Andrew McGovern, Jeremy van Vlymen, Ivelina Yonova, Simon Jones, and Simon de Lusignan. Royal College of General Practitioners Research and Surveillance Centre (RCGP RSC) sentinel network: A cohort profile. *BMJ open*, 6(4):e011092, April 2016.

[29] H. Zhao, H. Green, A. Lackenby, M. Donati, J. Ellis, C. Thompson, A. Bermingham, J. Field, P. Sebastianpillai, M. Zambon, Jm Watson, and R. Pebody. A new laboratory-based surveillance system (Respiratory DataMart System) for influenza and other respiratory viruses in England: Results and experience from 2009 to 2012. *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, 19(3), January 2014.

[30] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, Janneke Heijne, Malgorzata Sadkowska-Todys, Magdalena Rosinska, and W. John Edmunds. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS medicine*, 5(3):e74, March 2008.

[31] Helen J Wearing, Pejman Rohani, and Matt J Keeling. Appropriate Models for the Management of Infectious Diseases. *PLoS Medicine*, 2(7), July 2005.

[32] O. Diekmann, J. a. P. Heesterbeek, and M. G. Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society, Interface*, 7(47):873–885, June 2010.

[33] Fabrice Carrat, Elisabeta Vergu, Neil M. Ferguson, Magali Lemaitre, Simon Cauchemez, Steve Leach, and Alain-Jacques Valleron. Time lines of infection and disease in human influenza: A review of volunteer challenge studies. *American Journal of Epidemiology*, 167(7):775–785, April 2008.

[34] Douglas M. Fleming, Robert J. Taylor, Roger L. Lustig, Cynthia Schuck-Paim, François Haguinet, David J. Webb, John Logie, Gonçalo Matias, and Sylvia Taylor. Modelling estimates of the burden of Respiratory Syncytial virus infection in adults and the elderly in the United Kingdom. *BMC infectious diseases*, 15:443, October 2015.

[35] R. J. Pitman, L. J. White, and M. Sculpher. Estimating the clinical impact of introducing paediatric influenza vaccination in England and Wales. *Vaccine*, 30(6):1208–1224, February 2012.

[36] Health Protection Agency. Surveillance of influenza and other respiratory viruses in the UK (2010-2011 report). http://webarchive.nationalarchives.gov.uk/20140629102627/http:/hpa.org.uk/webc/HPAwebFile/HPAweb_C/1296687414154, 2011.

[37] Health Protection Agency. Surveillance of influenza and other respiratory pathogens in the UK 2012. http://webarchive.nationalarchives.gov.uk/20140714084352/http:/www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1317134705939, 2012.

[38] Public Health England. Surveillance of influenza and other respiratory viruses, including novel respiratory viruses, in the United Kingdom: Winter 2012/13. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/325217/Annual_flu_report_winter_2012_to_2013.pdf, 2013.

[39] Public Health England. Surveillance of influenza and other respiratory viruses in the United Kingdom: Winter 2013/14. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/325203/Flu_annual_report_June_2014.pdf, 2014.

[40] R. G. Pebody, H. K. Green, N. Andrews, H. Zhao, N. Boddington, Z. Bawa, H. Durnall, N. Singh, A. Sunderland, L. Letley, J. Ellis, A. J. Elliot, M. Donati, G. E. Smith, S. de Lusignan, and M. Zambon. Uptake and impact of a new live attenuated influenza vaccine programme in England: Early results of a pilot in primary school-age children, 2013/14 influenza season. *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, 19(22), June 2014.

[41] Lisa A. Grohskopf, Leslie Z. Sokolow, Karen R. Broder, Sonja J. Olsen, Ruth A. Karron, Daniel B. Jernigan, and Joseph S. Bresee. Prevention and Control of Seasonal Influenza with Vaccines. *MMWR. Recommendations and reports: Morbidity and mortality weekly report. Recommendations and reports*, 65(5):1–54, August 2016.

[42] Pia Hardelid, Greta Rait, Ruth Gilbert, and Irene Petersen. Recording of Influenza-Like Illness in UK Primary Care 1995-2013: Cohort Study. *PloS One*, 10(9):e0138659, 2015.

Chapter 5

Discussion

The objective of this thesis was to apply Bayesian methods to problems in network inference, evidence synthesis for risk assessment and infectious disease modelling. While the three problems differed in research questions and methodologies, in all three cases we sought to create models that are methodologically novel, fit-for-purpose but also generalisable. In one case the new method was implemented as part of new statistical software, while in the case of meta-analyses we provided generic and reproducible models. In all three cases, while the immediate objectives were achieved, new potential directions of inquiry were identified during and after the research. We try to avoid restating contents of discussion and conclusion parts of the manuscripts included in Chapters 2, 3 and 4, but rather aim provide a short and practical proposal for further work in each of the three cases.

Bayesian networks

In modelling networks, a new algorithm was devised and implemented to allow for presence of cyclical structures in graphical models. In the presented method, graphs are scored by combining their marginalised likelihood and graph priors (including priors on occurrences of structures such as cycles), which means that the inference is fully Bayesian. Efficient implementation of network inference is possible via a Markov Chain Monte Carlo program coded in C, which has been modified to allow for modelling of cyclical structures through the use of condensed graphs. In the result, the method can efficiently output a full posterior distribution (over space of graphs), in contrast to typical approaches. The output of the work on Bayesian networks comprises a paper, open-source software code in C, and a further software package in R.

In the next step, the new approach to modelling cyclic structures should be tested on non-linear data, which are common in biological networks. This work should proceed in stages, starting with linear approximations of non-linear phenomena but ultimately non-linear models of relationships between nodes may be required. Some part of these experiments should involve working with longitudinal data. Currently the method offers computational benefits, and can be useful as an alternative means of describing joint probability distributions through the use of multivariate distributions alongside univariate ones, but both methodological work and case studies will be needed to better evaluate its ability to learn cyclic structures in causal graphs in real settings. The one large scale case study on synthetic interventional data in gene regulatory network which we conducted with *graph_sampler* was promising, but our case study was performed on an acyclic causal graph.

Bayesian meta-analysis in risk assessment

A novel meta-analysis model was used to characterise differences in toxicokinetic parameters across subgroups of the population. Such characterisation is important from the viewpoint of chemical risk assessment, where safe levels of exposure to chemicals can be better informed through an understanding of intraindividual variability. The proposed model addresses issues with input data that are typical in risk assessment and offers three modelling advantages over simple meta-analysis models. One, accounting for variability in means, including the impact of subgroups (in case of polymorphic enzymes this is done by introducing priors and some biologically plausible constraints). Two, simultaneous modelling of means and variances. Three, a way to combine individual-level and study-level data. The output of this work comprises a modelling paper, including generic modelling codes and an application paper in modelling of mixtures of compounds. Further work should concentrate on applying the presented model to existing datasets. At the time of writing, we are using the model for two additional analyses of toxicokinetics of enzymes associated with CYP2C9 and CYP2C19 genes, which we hope to publish jointly. Peer review of these analyses will offer us a chance to improve the core model and in particular test the cross-validation approach, which was not used in model building for the mixtures of chemicals paper. However, regardless of work on applications, a few other methodological questions should be resolved separately. In particular, we would like to return to the question of the impact of different measures of dispersion and small sample sizes on the final result. As suggested in our work, this should involve both a simulation study approach (to understand model performance) and a modification of the model to allow for measurement error (e.g. through a generalised gamma distribution).

Influenza modelling

In the study of influenza vaccination, a Bayesian model was used to understand the impact of counterfactual health policy scenario of vaccinating children in the United Kingdom. To this end, parameters of both the epidemiological model (of the spread of influenza) and surveillance model (linking latent influenza infections to the observed healthcare events of different medical diagnoses, virus testing, hospitalisations, deaths) were estimated jointly. Using the inferred parameter distributions, counterfactual scenarios were simulated across 4 different influenza seasons, providing a measure of the impact of an alternative vaccination policy which accounted for model uncertainty. A statistical modelling paper based on this work has been published, together with additional observational, retrospective study concerning the influenza epidemiology.

The model presented in this thesis was devised with a specific study objective and a specific set of data in mind. However, no primary data collection was required and the data required for replicating this approach are routinely available in real-world observational databases and even in publicly available data sets that already exist in some countries. Therefore the model could be adapted to other countries, other seasons or even to perform within-season forecasting of epidemics, without major modifications to the likelihood function. However, to make the model adaptable and reproducible, additional work will be required. As the first step we propose deriving a generic version of the MCMC sampler which could work with different types of influenza surveillance data. As a default we suggest implementation in Stan, which allows for fast inference on non-linear models and would make this approach reproducible. Additionally, we are not aware of any existing generic solutions for compartmental models of infectious diseases in Stan, so such a contribution would be to the research community at large.

Bibliography

Alon, U. (2007). Network motifs: Theory and experimental approaches. *Nature Reviews Genetics*, 8:450–461. 13

Baguelin, M., Flasche, S., Camacho, A., Demiris, N., Miller, E., and Edmunds, W. J. (08-Oct-2013). Assessing Optimal Target Populations for Influenza Vaccination Programmes: An Evidence Synthesis and Modelling Study. *PLOS Medicine*, 10(10):e1001527. 00155. 19

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Molecular Systems Biology*, 3(1):78. 12

Blair, R. H., Kliebenstein, D. J., and Churchill, G. A. (05-Apr-2012). What Can Causal Networks Tell Us about Metabolic Pathways? *PLOS Computational Biology*, 8(4):e1002458. 12

Bois, F. Y. and Gayraud, G. (2015). Probabilistic generation of random networks taking into account information on motifs occurrence. *Journal of Computational Biology*, 22(1):25–36. 13

Chao, D. L., Halloran, M. E., Obenchain, V. J., and Jr, I. M. L. (29-Jan-2010). FluTE, a Publicly Available Stochastic Influenza Epidemic Simulation Model. *PLOS Computational Biology*, 6(1):e1000656. 17

Chickering, D. M. (2002). Optimal Structure Identification With Greedy Search. *Journal of Machine Learning Research*, 3(Nov):507–554. 13

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298. 00665. 9

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, Oxford, New York. 00000. 9

Cohen, J. (1994). The earth is round ($p < .05$). 10

Datta, S., Gayraud, G., Leclerc, E., and Bois, F. Y. (2017). Graph_sampler: A simple tool for fully Bayesian analyses of DAG-models. *Computational Statistics*, 32(2):691–716. 13

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, 7(3-4):601–620. 12

Gani, R., Hughes, H., Fleming, D., Griffin, T., Medlock, J., and Leach, S. (2005). Potential Impact of Antiviral Drug Use during Influenza Pandemic. *Emerging Infectious Diseases*, 11(9):1355–1362. 18

Geiger, D. and Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, 30(5):1412–1440. 13

Gelman, A. (2005). Analysis of variance?why it is more important than ever. *The Annals of Statistics*, 33(1):1–53. 00000. 14

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534. 00000. 14

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press. 10

Gelman, A. and Robert, C. P. (2010). "Not only defended but also applied": The perceived absurdity of Bayesian inference. 00000. 10

Gilbert, N. (2007). *Agent-Based Models*. SAGE Publications, Inc, Los Angeles, 1 edition edition. 17

Haavelmo, T. (1943). The Statistical Implications of a System of Simultaneous Equations. *Econometrica*, 11(1):1–12. 12

Hardy, A., Benford, D., Halldorsson, T., Jeger, M. J., Knutsen, H. K., More, S., Naegeli, H., Noteborn, H., Ockleford, C., Ricci, A., Rychen, G., Schlatter, J. R., Silano, V., Solecki, R., Turck, D., Benfenati, E., Chaudhry, Q. M., Craig, P., Frampton, G., Greiner, M., Hart, A., Hogstrand, C., Lambre, C., Luttik, R., Makowski, D., Siani, A., Wahlstroem, H., Aguilera, J., Dorne, J.-L., Fernandez Dumont, A., Hempen, M., Valtueña Martínez, S., Martino, L., Smeraldi, C., Terron, A., Georgiadis, N., and Younes, M. (2017a). Guidance on the use of the weight of evidence approach in scientific assessments. *EFSA Journal*, 15(8):e04971. 16

Hardy, A., Benford, D., Halldorsson, T., Jeger, M. J., Knutsen, K. H., More, S., Mortensen, A., Naegeli, H., Noteborn, H., Ockleford, C., Ricci, A., Rychen, G., Silano, V., Solecki, R., Turck, D., Aerts, M., Bodin, L., Davis, A., Edler, L., Gundert-Remy, U., Sand, S., Slob, W., Bottex, B., Abrahantes, J. C., Marques, D. C., Kass, G., and Schlatter, J. R. (2017b). Update: Use of the benchmark dose approach in risk assessment. *EFSA Journal*, 15(1):e04658. 15

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. page 31. 9

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics (Oxford, England)*, 19(17):2271–2282. 12

Ioannidis, J. P. A. (2018). The Proposal to Lower P Value Thresholds to .005. *JAMA*, 319(14):1429–1430. 00000. 10

Ioannidis, J. P. A. (2019). What Have We (Not) Learnt from Millions of Scientific Papers with P Values? *The American Statistician*, 73(sup1):20–25. 10

Ioannidis, J. P. A. (30-Aug-2005). Why Most Published Research Findings Are False. *PLOS Medicine*, 2(8):e124. 10

Jansen, J. P. (2011). Network meta-analysis of survival data with fractional polynomials. *BMC Medical Research Methodology*, 11:61. 15

Kamal, M. A., Smith, P. F., Chaiyakunapruk, N., Wu, D. B. C., Pratoomsoot, C., Lee, K. K. C., Chong, H. Y., Nelson, R. E., Nieforth, K., Dall, G., Toovey, S., Kong, D. C. M., Kamauu, A., Kirkpatrick, C. M., and Rayner, C. R. (2017). Interdisciplinary pharmacometrics linking oseltamivir pharmacology, influenza epidemiology and health economics to inform antiviral use in pandemics: Linking pharmacology to influenza epidemiology and health economics. *British Journal of Clinical Pharmacology*, 83(7):1580–1594. 18

Koster, J. T. A. (1996). Markov properties of nonrecursive causal models. *The Annals of Statistics*, 24(5):2148–2177. 13

Kruschke, J. K. and Vanpaemel, W. (2015). *Bayesian Estimation in Hierarchical Models*, volume 1. Oxford University Press. 00000. 10

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067. 9

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press. 00000. 10

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician*, 73(sup1):235–245. 00000. 10

Meager, R. (2019). Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments. *American Economic Journal: Applied Economics*, 11(1):57–91. 00000. 15

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., and Edmunds, W. J. (2008). Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. *PLOS Medicine*, 5(3):e74. 18

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer New York, New York, NY. 9

NHS (2017). Children's flu vaccine. https://www.nhs.uk/conditions/vaccinations/child-flu-vaccine/. 20

Osthus, D., Gattiker, J., Priedhorsky, R., and Del Valle, S. Y. (2017). Dynamic Bayesian Influenza Forecasting in the United States with Hierarchical Discrepancy. *arXiv:1708.09481 [stat]*. 19

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition. 00000. 9, 12

Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. (2013). Causal discovery with continuous additive noise models. *arXiv:1309.6779 [stat]*. 13

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. page 10. 9

Quignot, N., Wiecek, W., Amzal, B., and Dorne, J.-L. (2018). The Yin–Yang of CYP3A4: A Bayesian meta-analysis to quantify inhibition and induction of CYP3A4 metabolism in humans and refine uncertainty factors for mixture risk assessment. *Archives of Toxicology*. 47

Rajaram, S., Wiecek, W., Lawson, R., Blak, B., Zhao, Y., Hackett, J., Brody, R., Salimi, T., Amzal, B., and Patel, V. (2018). A retrospective observational analysis of post-pandemic influenza-related outcomes in the United Kingdom, 2010-2014. *Human Vaccines & Immunotherapeutics*, 14(2):368–377. 87

Rajaram, S., Wiecek, W., Lawson, R., Blak, B. T., Zhao, Y., Hackett, J., Brody, R., Patel, V., and Amzal, B. (2017). Impact of increased influenza vaccination in 2-3-year-old children on disease burden within the general population: A Bayesian model-based approach. *PloS One*, 12(12):e0186739. 87

Renwick, A. G. and Lazarus, N. R. (1998). Human Variability and Noncancer Risk Assessment— An Analysis of the Default Uncertainty Factor. *Regulatory Toxicology and Pharmacology*, 27(1):3–20. 15

Richardson, T. (1996). A Discovery Algorithm for Directed Cyclic Graphs. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, pages 454–461, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 13

Robins, J., Hernán, M., and Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, 11(5):550–560. 12

Rubin, D. B. (1981). Estimation in Parallel Randomized Experiments. *Journal of Educational Statistics*, 6(4):377–401. 14

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331. 12

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, 308(5721):523–529. 12

Simonsen, L., Clarke, M. J., Schonberger, L. B., Arden, N. H., Cox, N. J., and Fukuda, K. (1998). Pandemic versus Epidemic Influenza Mortality: A Pattern of Changing Age Distribution. *The Journal of Infectious Diseases*, 178(1):53–60. 19

Simonsen, L., Reichert, T. A., Viboud, C., Blackwelder, W. C., Taylor, R. J., and Miller, M. A. (2005). Impact of Influenza Vaccination on Seasonal Mortality in the US Elderly Population. *Archives of Internal Medicine*, 165(3):265–272. 19

Spirtes, P. (1995). Directed Cyclic Graphical Representations of Feedback Models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 491–498, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 13

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. Lecture Notes in Statistics. Springer-Verlag, New York. 13

van Leeuwen, E., Klepac, P., Thorrington, D., Pebody, R., and Baguelin, M. (20-Nov-2017). fluEvidenceSynthesis: An R package for evidence synthesis based analysis of epidemiological outbreaks. *PLOS Computational Biology*, 13(11):e1005838. 19

Vardavas, R., Breban, R., and Blower, S. (04-May-2007). Can Influenza Epidemics Be Prevented by Voluntary Vaccination? *PLOS Computational Biology*, 3(5):e85. 18

Vidgen, B. and Yasseri, T. (2016). P-Values: Misunderstood and Misused. *Frontiers in Physics*, 4. 10

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133. 00000. 10

Wearing, H. J., Rohani, P., and Keeling, M. J. (2005). Appropriate Models for the Management of Infectious Diseases. *PLoS Medicine*, 2(7). 17

Weber, S., Gelman, A., Lee, D., Betancourt, M., Vehtari, A., and Racine-Poon, A. (2018). Bayesian aggregation of average data: An application in drug development. *The Annals of Applied Statistics*, 12(3):1583–1604. 00002. 15

Werhli, A. V. and Husmeier, D. (2007). Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge. *Statistical Applications in Genetics and Molecular Biology*, 6(1). 12

Wiecek, W., Amzal, B., Bakshi, S., Patel, V., and Staa, T. V. (2015). Age Related Consultation Rates of Clinically-Diagnosed Influenza And Acute Respiratory Illnesses Observed Through A Network of Gp Practices Across England. *Value in Health*, 18(7):A579. 87

Wiecek, W., Bois, F., and Gayraud, G. (2019a). Structure learning of Bayesian networks involving cyclic structures. 21

Wiecek, W., Dorne, J.-L., Quignot, N., Bechaux, C., and Amzal, B. (2019b). A generic Bayesian hierarchical model for the meta-analysis of human population variability in kinetics and its applications in chemical risk assessment. *Computational Toxicology*, page 100106. 47

Wright, S. (1921). Correlation and causation. *Journal of agricultural research*, 20(7):557–585. 12